Frequentist and Bayesian analysis methods for case series data and application to early outpatient COVID-19 treatment case series

Eleftherios Gkioulekas, Ph.D.

Professor, School of Mathematical and Statistical Sciences, University of Texas Rio Grande Valley, Edinburg, TX, United States[*]

Peter A McCullough, MD, MPH

Chief Medical Advisor, Truth for Health Foundation, Tucson AZ, United States[†]

Vladimir Zelenko, MD

Affiliate Physician, Columbia University Irving Medical Center, New York City, NY, United States[‡]

When confronted with a public health emergency, significant innovative treatment protocols can sometimes be discovered by medical doctors at the front lines based on repurposed medications. We propose a very simple hybrid statistical framework for analyzing the case series of patients treated with such new protocols, that enables a comparison with our prior knowledge of expected outcomes, in the absence of treatment. The goal of the proposed methodology is not to provide a precise measurement of treatment efficacy, but to establish the existence of treatment efficacy, in order to facilitate the binary decision of whether the treatment protocol should be adopted on an emergency basis. The methodology consists of a frequentist component that compares a treatment group against the unknown probability of an adverse outcome in the absence of treatment, and calculates a lower bound for this unknown probability, that has to be exceeded, in order to control the corresponding p-value, and reject the null hypothesis. We explain the relationship of this method with the exact Fisher test and the binomial proportion confidence interval problem. The resulting lower bound (hereafter, efficacy threshold) is further adjusted with a Bayesian technique, in order to also control the false positive rate. The combined techniques are applied to case series of high-risk COVID-19 outpatients, that were treated using the early Zelenko protocol and the more enhanced McCullough protocol. The resulting efficacy thresholds are then compared against our prior knowledge of mortality and hospitalization rates of high-risk COVID-19 patients, as reported in the research literature.

Keywords: COVID-19; SARS-CoV-2; ambulatory treatment; early treatment, mortality; hospitalization; epidemiology; biostatistics; drug repurposing.

1. INTRODUCTION

In medical research, the efficacy of new drugs or treatment protocols is established by controlled studies in which a treatment group is compared against a control group. A case series is one half of a controlled study consisting only of the treatment group. At the beginning of the COVID-19 pandemic, practicing medical doctors were confronted with having no treatment to offer to their patients that can prevent or minimize hospitalization and/or death. In response, some doctors were compelled to innovate and discover, on their own, treatment protocols using repurposed off-label medications. Most notable examples, amongst several others, include Didier Raoult [1] in the IHU Méditerranée Infection hospital in Marseilles France, Vladimir Zelenko [2] in upstate New York, Shankara Chetty [3] in South Africa, and Paul Marik's group [4], which was in the beginning based at the Eastern Virginia Medical School. Their efforts to treat patients generated case series of successfully treated patients that constitute Real-World Evidence [5].

The goal of this paper is to present a statistical framework for analyzing systematic case series data of early treatment protocols with binary endpoints (*e.g.* hospitalization or death), and comparing them against our prior knowledge of the likelihood of adverse outcomes in the absence of treatment. Under certain conditions, which we shall elaborate on below, the proposed methodology can be used to establish the existence of treatment efficacy, but falls short of precisely measuring the corresponding odds ratio. Nevertheless, this can be enough evidence to justify a positive recommendation to adopt these treatment protocols, on an emergency basis, while more detailed clinical research is in progress. Following the recommendation of the American Statistical Association statement on statistical significance and p-values [6], the proposed approach combines use of the p-value, which enables one to reject the null hypothesis, with a Bayesian factor analysis framework [7–11] for controlling the false positive rate [12]. Empirically, we have found that the frequentist p-value framework has done a pretty good job on its own, at least for the analysis of the case series data considered in this paper. However, complementing it with Bayesian factor analysis can help raise the red flag when dealing with small sample sizes and/or weak

[*]Electronic address: drlf@hushmail.com
[†]Electronic address: peteramccullough@gmail.com
[‡]Electronic address: zz613@hotmail.com

signals.

We apply the proposed framework to the processing of available case series data [2, 13–17] that support proposed early outpatient treatment protocols for COVID-19 patients, such as the original Zelenko triple-drug protocol [2] and the more advanced McCullough protocol [18–20]. The original Zelenko protocol was first announced on March 23, 2020 [21]. The proposed approach was to risk-stratify patients into two groups (low-risk vs high-risk), provide supportive care to the low-risk group, and treat the high-risk group with a triple-drug protocol (hydroxychloroquine, azithromycin, zinc sulfate). Zelenko's risk stratification criteria included in the high-risk group the following: all patients older than 60 were classified as high-risk; all patients of any age with at least one comorbidity or BMI > 30 kg/m^2; all patients of any age that presented with shortness of breath. Results were reported in an April 28, 2020 letter [13] and a June 14, 2020 letter [14], and the lab-confirmed subset of the April data was published in a formal case-control study [2]. Zelenko's letters have been attached to our supplementary material document [22].

The rationale for the triple-drug therapy was based on the following mechanisms of action: Hydroxychloroquine prevents the virus from binding with the cells, and also acts as a zinc ionophore that brings the zinc ions inside the cells, which in turn inhibit the RDRP (RNA Dependent RNA Polymerase) enzyme used by the virus to replicate [23, 24]. Azithromycin's role is to guard against a secondary infection, but we have since learned that it also has it's own anti-viral properties [25–27], and a signal of the efficacy of adding azithromycin on top of hydroxychloroquine can be clearly discerned in a study of nursing home patients in Andorra, Spain [28].

It is interesting that chloroquine was shown *in vitro* to have antiviral properties against the previous SARS-CoV-1 virus [29], and that there is an anecdotal report from 1918 [30] about the successful use of quinine dihydrochloride injections as an early treatment of the Spanish flu. In hindsight, it is now known that influenza viruses also use the RDRP protein to replicate [31], which can be inhibited with intracellural zinc ions [23, 24]. Consequently, there is a mechanism of action that can explain why we should anticipate the combination of zinc with a zinc ionophore (i.e. hydroxychloroquine, or quercetin [32], or EGCG [33]) to inhibit the replication of the influenza viruses. Other RNA viruses, including the respiratory syncytial virus (RSV) [34] and the highly pathogenic Marburg and Ebola viruses [35, 36], are also using the RDRP protein to replicate. It remains a compelling hypothesis, deserving further investigation, that the zinc/zinc ionophore concept could play an important role as part of a broader multi-drug treatment or prophylaxis protocol against these serious infectious diseases.

Zelenko's protocol was soon extended into a sequenced multi-drug approach, known as the McCullough protocol [18–20], which is based on the insight that COVID-19 is a tri-phasic illness that manifests in three phases: (1) an initial viral replication phase, in which the virus infects cells and uses them to replicate and make new viral particles, during which patients present with flu-like symptoms; (2) an inflammatory hyper-dysregulated immune-modulatory florid pneumonia, that presents with a cytokine storm, coughing, and shortness of breath, triggered by the toxicity of the spike protein [37], when it is released, as viral particles are destroyed by the immune system, triggering release of interleukin-6 and a wave of cytokines; (3) a thromboembolic phase, during which microscopic blood clots develop in the lungs and the vascular system, causing oxygen desaturation, and very damaging complications that can include embolic stroke, deep vein thrombosis, pulmonary embolism, myocardial injury, heart attacks, and damage to other organs.

The rationale of the original Zelenko protocol was that early intervention to stop the initial viral replication phase could prevent the disease from progressing to the second and third phase, and, in doing so, prevent hospitalizations or death. The McCullough protocol [18–20] extends the Zelenko protocol by using multiple drugs in combination sequentially to mitigate each of the three phases of the illness, depending on how they present for each individual patient. McCullough's therapeutic recommendations for handling the cytokine injury phase and the thrombosis phase of the COVID-19 illness are, for the most part, standard on-label treatments for treating hyper-inflammation and preventing blood clots. The most noteworthy innovations to the antiviral part of the protocol are the addition of ivermectin, which has 20 mechanisms of action against COVID-19 [38], as an antiviral medication [39–44] to be used as an alternative or in conjunction with hydroxychloroquine, the addition of a nutraceutical bundle [45–47] combined with a zinc ionophore (quercetin [32] or EGCG [33]) for both low-risk and high-risk patients, and lowering the age threshold for high-risk patients to 50 years. The MATH+ protocol [4], developed for hospitalized patients by Marik's group, follows the same principles of a sequenced multi-drug treatment. A similar treatment protocol, based on similar insights, was independently discovered and published on May 2020 by Chetty [3] in South Africa.

McCullough's protocol [18–20] was adopted by some treatment centers throughout the United States and overseas, but has not been endorsed by the United States public health agencies, ostensibly due to lack of support of the entire sequenced treatment algorithm by a randomized controlled trial (hereafter RCT). In spite of the urgent need for safe and effective early outpatient treatment protocols for COVID-19, there has been no attempt to conduct any such trials of any comprehensive multi-drug outpatient treatment protocols throughout the pandemic. Instead, the prevailing approach has been to try to build treatment protocols, one drug at a time, after validating each drug with an RCT. Because COVID-19 is a multifaceted tri-phasic illness, and there is no a priori reason to expect that a single drug alone will work for all 3 phases of the disease, this orthodox approach is not an optimal research strategy. The first priority should be to validate the efficacy of treatment protocols that use multiple drugs in combination, since this is what is actually going to be used in practice to treat patients. On that front, there are published observational data [15–17] on the efficacy of the McCullough protocol, in addition to Zelenko's reported outcomes [2, 13, 14, 21], and there is an abundance of unpublished real world data from several treatment centers, from around the world, that have yet to

be analyzed [48]. The statistical framework proposed in this paper is the missing link for conducting a formal analysis of the available observational data on early COVID-19 treatment protocols.

The broader context in which the proposed statistical methodology is situated is as follows. Shortly before COVID-19 was declared a pandemic by the World Health Organization, an article [49] was published on February 23, 2020 in the New England Journal of Medicine arguing that *"the replacement of randomized trials with non-randomized observational status is a false solution to the serious problem of ensuring that patients receive treatments that are both safe and effective"*. The opposing viewpoint was published earlier in 2017 by Frieden [50], highlighting the limitations of RCTs and the need to leverage and overcome the limitations of all available sources of evidence, including real world evidence [5], in order to make lifesaving public health decisions. In particular, Frieden [50] stressed that the very high cost of RCTs and the long timelines needed for planning, recruiting patients, conducting the study, and publishing it, are limitations that *"affect the use of randomized controlled trials for urgent health issues, such as infectious disease outbreaks for which public health decisions must be made quickly on the basis of limited and imperfect data."*

Deaton and Cartwright [51] presented the conceptual framework that underlies RCTs and highlighted several limitations. Among them, they have stressed that randomization requires very large samples on both arms of the trial, otherwise, an RCT should not be presumed to be methodologically superior to a corresponding observational study. Furthermore, although a properly conducted RCT has internal validity, in that the inferences are applicable to the specific group of patients that participated in the trial, the external validity of the RCT outcomes needs to be justified conceptually on the basis of prior knowledge, which is either observational, or based on a deeper understanding of the underlying mechanisms of action. Because COVID-19 mortality risk in the absence of early treatment can span three orders of magnitude (from 0.01% to more than 10%) [52–58], depending on age and comorbidities, trials using low-risk patient cohorts are not informative about expected outcomes on the high-risk patient cohorts and vice versa.

Furthermore, as was noted by Risch [59], when interpreting evidence from RCTs, and more broadly from any study, we should bear in mind that results of efficacy or toxicity of a treatment regimen on hospitalized patients cannot be extrapolated to outpatients and vice versa. Likewise, Risch [59] noted that evidence of efficacy or lack of efficacy of a single drug do not necessarily extrapolate to using several drugs in combination. This latter point is further amplified when there is an algorithmic overlay governing, which drugs should be used and when, based on the individual patient's medical history and ongoing response to treatment.

In addition to all that, we are also confronted with an ethical concern. If the available observational evidence are sufficiently convincing, then there is a crossover point where it is no longer ethical to justify randomly refusing treatment to a large number of patients, in order to have a sufficiently large control group. The corresponding mathematical challenge is being able to quantify the quality of our observational evidence in order to determine whether or not we are already situated beyond this ethical crossover point.

Just as the quality of evidence provided by randomized controlled trials is fluid with respect to successful randomization and external validity, the same is true for the real world evidence [5] obtained from any uncontrolled case series. Although lacking a control prevents us from measuring the corresponding odds ratio, the confluence of the following conditions makes it possible to establish the existence of treatment efficacy: First, the proposed treatment protocols should use repurposed drugs [60] with a known excellent safety record. When testing new drugs, we have no prior knowledge of the risks involved and a rigorous controlled study is required to determine the balance of risks and benefits. Both Zelenko's triple drug therapy [2] as well as the expanded multi-drug McCullough protocol [18–20] for the early outpatient treatment of COVID-19 are based exclusively on safe repurposed medications. Second, we need data that give us prior knowledge of the probability risk of the relevant binary endpoints (*i.e.* hospitalization and/or death) in the absence of treatment, as a function of the relevant stratification parameters. It is not necessary to have a comprehensive model and it may just be sufficient to be able to obtain a good lower bound for the respective probability risk, in the absence of treatment. Third, and most importantly, the case series corresponding to treated patients should exhibit a *very strong* signal of benefit, relative to our prior experience with untreated patients, prior to the discovery of the respective treatment protocol.

In simple operational terms, the idea that is proposed in this paper works as follows. Our input is the number N of high-risk patients treated, the number of patients a with an adverse outcome (i.e. hospitalization or death) and selection criteria for extracting the high-risk cohort under consideration, from which we can deduce, based on prior knowledge, that the unknown probability x of an adverse outcome without treatment is bounded by $p_2 > x > p_1$. We also choose the desired level of p-value upper bound p_0, which is typically $p_0 = 0.05$ (95% confidence), although we shall also consider $p_0 = 0.01$ and $p_0 = 0.001$. The output is an efficacy threshold $x_0^+(N, a, p_0)$ that gives us the following rigorous mathematical statement: *if* $x > x_0^+(N, a, p_0)$*, then we have more than* $1 - p_0$ *confidence that the treatment is effective relative to the standard of care.* This statement has to be paired with the subjective assessment of our prior knowledge, based on which we need to show that $p_1 > x_0^+(N, a, p_0)$. The upper bound p_2 is used by the Bayesian factor technique as part of finalizing the calculation of the efficacy threshold $x_0^+(N, a, p_0)$. When there is a large gap between p_1 and $x_0^+(N, a, p_0)$, and furthermore, when the treatment relies on repurposed drugs with known excellent safety record, then we have clear and convincing evidence that the treatment is effective. On the other hand, when new medications, as opposed to repurposed drugs, are introduced into a preexisting treatment protocol, they should be rigorously studied both for safety and efficacy with prospective RCTs.

The paper is organized as follows. On Section 2 we present the technique for calculating the efficacy threshold

$x_0^+(N, a, p_0)$ that needs to be exceeded by the probability x of an adverse event without treatment, in order to reject the null hypothesis and control the corresponding p-value. We also explain the relationship of the proposed technique with the exact Fisher test and with the binomial proportion confidence interval problem. On Section 3, we present a Bayesian technique for adjusting the efficacy thresholds $x_0^+(N, a, p_0)$ in order to also control the corresponding false positive rate. In Section 4, we illustrate an application of both techniques to the Zelenko case series [2, 13, 14] as well as the Procter [15, 16] and Raoult [17] case series. Discussion and conclusions are given in Section 5.

2. FREQUENTIST METHODS FOR CASE SERIES ANALYSIS

In this section, we present the technique for comparing a treatment group case series against the expected probability x of an adverse outcome without treatment, based on prior knowledge. Since the probability x is unknown, we calculate the minimum value that this probability has to exceed in order to be able to reject the null hypothesis, that the treatment has no efficacy. The proposed technique is equivalent to an exact Fisher test where we take the limit of an infinitely large control group with probability of an adverse outcome set equal to x. Contrary to what one might expect, although this limit does converge, it does not do so monotonically. Likewise, when comparing a treatment group (N, a) with the probability x of an adverse outcome without treatment, we find that the resulting p-value is not monotonic with respect to x either. We also explain the relationship of the proposed approach with a binomial confidence interval problem, and provide evidence that the corresponding coverage probability is conservative.

2.1. Comparing treatment group against expected adverse event rate without treatment.

Suppose that we have a treatment group of patients in which N patients have received treatment, and a patients have had an adverse outcome. Without an appropriate control group, there is no way to determine what would have happened to that same group of patients if they had not been treated. Let x be this unknown probability of an adverse outcome without treatment. Although x may be unknown, we can nevertheless calculate a lower bound $x_0^+(N, a, p_0)$ such that if $x > x_0^+(N, a, p_0)$ then the p-value $p(N, a, x)$, corresponding to observing the event (N, a) under the null hypothesis, satisfies $p(N, a, x) < p_0$, with $p_0 = 0.05$ for 95% confidence in rejecting the null hypothesis, or alternatively $p_0 = 0.01$ for 99% confidence, and $p_0 = 0.001$ for 99.9% confidence.

First, we note that under the null hypothesis, that the treatment applied to the treatment group is ineffective, the probability of observing a patients with an adverse outcome out of a total of N patients is given by

$$\mathrm{pr}(N, a|x) = \binom{N}{a} x^a (1-x)^{N-a}, \tag{1}$$

which corresponds to a binomial distribution. The first factor gives the number of combinations for choosing the a patients that have an adverse outcome out of all N patients. The second factor x^a is the probability that the chosen a patients have an adverse outcome, under the assumption of the null hypothesis. The third factor $(1-x)^{N-a}$ is likewise the probability that the remaining $N-a$ patients will not have an adverse outcome. Consequently, the product of the three factors is the probability of seeing the event (N, a) under the null hypothesis.

The corresponding p-value is calculated by adding to the probability of the event (N, a), the probability of all other events with smaller or equal probability, and it reads

$$p(N, a, x) = \sum_{n=0}^{N} \mathrm{pr}(N, n|x) H(\mathrm{pr}(N, a|x) - \mathrm{pr}(N, n|x)), \tag{2}$$

where H is the modified Heaviside function given by

$$H(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}. \tag{3}$$

Given the analytic equation for the p-value as a function of x, the formal definition for the efficacy threshold $x_0^+(N, a, p_0)$ is given by

$$x_0^+(N, a, p_0) = \inf\{x \in [a/N, 1] \mid p(N, a, x) \leq p_0\}, \tag{4}$$

and we expect to find $x_0^+(N, a, p_0) > a/N$. By definition, the meaning of this efficacy threshold is that if the probability x of an adverse outcome without treatment exceeds $x_0^+(N, a, p_0)$, then the corresponding p-value for rejecting the null hypothesis satisfies $p(N, a, x) \leq p_0$, which is considered statistically significant for $p_0 < 0.05$. It should be stressed that the efficacy threshold $x_0^+(N, a, p_0)$ may have to be further increased, in accordance to the Bayesian methods presented in Section 3, in order to also control the false positive rate.

The reason why in the above equation x is restricted in the interval $[a/N, 1]$ is because one can also calculate another threshold $x_0^-(N, a, p_0)$ given by

$$x_0^-(N, a, p_0) = \sup\{x \in [0, a/N] \mid p(N, a, x) \leq p_0\}, \tag{5}$$

such that $x < x_0^-(N, a, p_0)$ implies $p(N, a, x) \leq p_0$. The $x_0^-(N, a, p_0)$ threshold is not relevant from the standpoint of analyzing a treatment group case series. However, it could be used on a control group case series to obtain a lower bound on the probability of an adverse outcome in an untreated cohort of patients, in the context of a specific confidence p_0.

For the case of a treatment group case series, lacking an appropriate control group, it is still possible to calculate the efficacy threshold $x_0^+(N, a, p_0)$ and we may be able to use it to provide evidence that the treatment used has some efficacy. As was explained in the introduction to the paper, to construct a convincing argument, several conditions must be satisfied. First, the medications used should have a known excellent safety profile so that any risk of an adverse outcome caused by the treatment itself, should be negligible relative to the risk posed by the disease itself, under the preexisting standard of

care. This requirement underscores that the proposed methodology should be used only when repurposing old medications [60] to address a new challenge. Second, our prior experience with the standard of care should be sufficiently detailed to be able to construct a predictive model of the risk of adverse outcome under the standard of care as a function of the demographic variables that are most closely associated with an increased risk. The risk model should provide a convincing estimate of the lower bound and upper bound of the probability of an adverse outcome in a demographic similar to that of the treatment group, when using the standard of care. Given such an interval (p_1, p_2) for the risk associated with the standard of care, we have statistical evidence in support of efficacy of the proposed treatment protocol if $x_0^+(N, a, p_0) < p_1 < p_2$. Third, there should be a substantial gap between the observed adverse outcome rate, a/N in the treatment group and the standard of care risk interval (p_1, p_2) in order to be able to plausibly rule out the placebo effect. Ideally, the gap between the efficacy threshold $x_0^+(N, a, p_0)$ and the risk interval (p_1, p_2) should also be as large as possible, which can be achieved with increased sample sizes.

Under these conditions, a comparison of the efficacy threshold $x_0^+(N, a, p_0)$ calculated from the treatment group case series, against the inferred risk interval (p_1, p_2) associated with the standard of care, can be used to provide statistical evidence in support of the existence of efficacy for the newly proposed treatment. Although it will not be possible to measure the efficacy, proving the existence of efficacy can be sufficient for recommending the adoption of a new treatment on an emergency basis.

2.2. Mathematical comments on the proposed hypothesis testing technique

We now redirect our focus towards some interesting mathematical observations. Let $p(N, a, M, b)$ be the p-value obtained from a two-tail exact Fisher test, with (N, a) being the treatment group of N patients with a patients having an adverse outcome, and (M, b) the control group of M patients, with b patients having an adverse outcome. It is reasonable to anticipate that $p(N, a, M, b)$ should be related with $p(N, a, x)$, in the sense that if we run an exact Fisher test using a hypothetical control group (M, b), such that $x = b/M$, then taking the limit in which the size of the control group goes to infinity should give us convergence to $p(N, a, x)$. From this, we can infer a corresponding relationship between the hypergeometric distribution, used in the calculation of $p(N, a, M, b)$, and the binomial distribution, used in the calculation of $p(N, a, x)$. As it turns out, this is a known result [61, 62], and there is even a detailed mathematical study [63] bounding the corresponding rate of convergence as the size of the control group approaches infinity.

In order to state the relationship between the two probabilities in a precise manner, we begin by noting that in Eq. (A1), the variable M appears only at the top argument of two binomial coefficients, one at the numerator, and one at the denominator. It follows that, notwithstanding that the variables

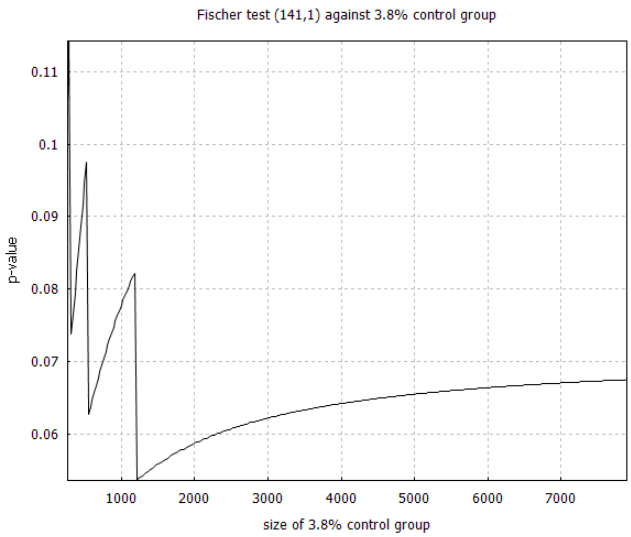

FIG. 1: We plot the p-value calculated from an exact Fisher test that compares the treatment group from the DSZ study [2] (141 high-risk patients treated with 1 death) against an artificial control group with 3.8% mortality rate. Note that the exact p-value in the infinite control group limit should be 0.047, which is approached to three decimals when we get to control group size between 160,000 and 180,000

N, a, M, b, are supposed to be integers, we can replace M with a continuous variable $(1/x)b$, take a discrete sequence limit $b \in \mathbb{N}^*$, with b going to infinity, and show that

$$\lim_{b \in \mathbb{N}^*} p(N, a, (1/x)b, b) = p(N, a|x). \tag{6}$$

We give a detailed proof of this equation in appendix A.

The paradoxical feature of the convergence of the exact Fisher test p-value, in the limit of an infinite control group, is that it does not converge monotonically. We illustrated this via an example in Fig. 1 and several more such example calculations have been included in our supplementary material document [22]. We have observed that as the size of the control group is increased, the p-value increases, which is counterintuitive, since we are expecting that a larger control group should increase the contrast between the treatment group and control group and thus decrease the p-value. On the other hand, the long term trend of the p-value is indeed that it tends to decrease, which is done by discontinuous downward jumps.

For an explanation of this phenomenon, we can surmise that as the hypergeometric distribution smoothly converges towards the binomial distribution, part of the curve is increasing, and another part is decreasing. The calculation shown in Fig. 1 implies that, amongst the probability terms contributing to the exact Fisher test p-value, the terms that tend to increase with increasing control group sample size, dominate over the terms that tend to decrease in the sum total. As they do so, eventually some term becomes larger than the probability term corresponding to the observed event (N, a, M, b), and is thus removed from the sum. The sudden removal of these terms explains the downward jumps and is the mechanism that actually drives the convergence of the exact Fisher p-value in the

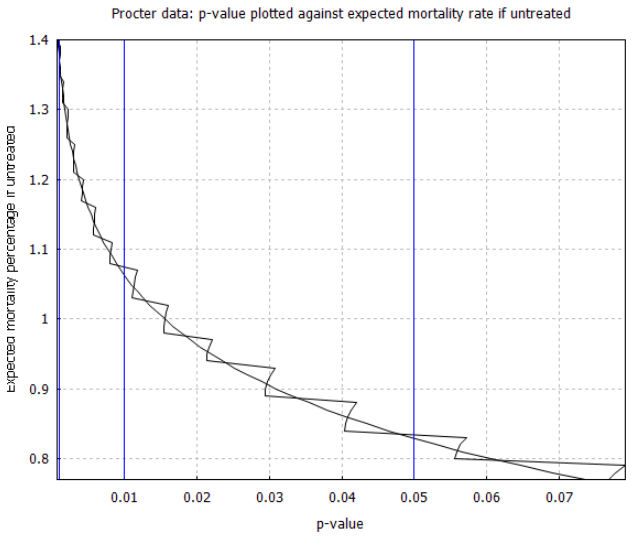

FIG. 2: Relationship between p-value and expected mortality rate for high risk patients without early treatment, based on the case series data from Procter's dataset of 869 high-risk patients [16]. The zigzag curve follows $p(N, a, x)$ given by Eq. (2), whereas the smooth curve follows $p_{\mathrm{CP}}(N, a, x)$ given by Eq. (7).

infinite control group limit.

We see similar behavior when plotting the p-value $p(N, a, x)$ as a function of x with constant N, a and increasing x. For example, in Fig. 2, we plot the p-value against the expected mortality rate without early outpatient treatment of COVID 19, based on Procter's combined case series [16] of 869 high risk patients with 2 deaths that received early treatment. The figure has vertical lines marking the crossover to 95%, 99%, and 99.9% confidence. With increasing x, the contrast between the mortality rate x that we are expecting without early treatment and the observed reduced mortality rate with early treatment also increases, and thus our confidence towards rejecting the null hypothesis should also improve. We are therefore expecting $p(N, a|x)$ to be monotonically decreasing with respect to x. Instead we see in Fig. 2 that $p(N, a|x)$ again follows a zigzag curve, where it increases with small increases in x, while exhibiting a decreasing trend with larger increases in x, that are driven by discontinuous downward jumps.

This behavior can be explained if we approximate $p(N, a, x)$ using the equation

$$p_{\mathrm{CP}}(N, a, x) = 2 \sum_{n=0}^{a} \mathrm{pr}(N, n|x), \tag{7}$$

which assumes, as an approximation, that the sum of the right-tail terms $\mathrm{pr}(N, n|x)$ with $a < n < N$, that are included in the exact p-value calculation, is equal to the corresponding left-tail sum. Replacing the right-tail sum with the left-tail sum gives us the smooth monotonic curve shown in Fig. 2. We can therefore identify the right-tail terms as the cause of the zigzag behavior in the exact p-value $p(N, a, x)$. We can also explain this theoretically by noting that the derivative of $\mathrm{pr}(N, n|x)$ with respect to x is given by,

$$(\mathrm{d}/\mathrm{d}x)\mathrm{pr}(N, n|x) = \binom{N}{n} (\mathrm{d}/\mathrm{d}x)[x^n (1-x)^{N-n}] \tag{8}$$

$$= \binom{N}{n} x^{n-1} (1-x)^{N-n-1} (n - Nx). \tag{9}$$

Under the assumption that we are looking at values of x that satisfy $x > a/N$, we see that for the left-tail terms we have $0 \le n < a$ and therefore $n - Nx < n - N(a/N) = n - a < 0$, and since all other factors in Eq. (9) are positive, we get the correct behavior that the left-tail terms are decreasing with respect to x. However, for the right-tail terms, we have $a < n \le N$ and therefore we have $n - Nx > 0$ if and only if n satisfies $a/N < x < n/N$. It follows that some, but not all of the terms in the right-tail sum, are going in the "wrong" direction and are increasing with respect to x. The right-tail terms that are decreasing are the ones that satisfy $a/N < n/N < x < 1$. The bad terms are eventually dropped from the p-value sum, when they become large enough, which accounts for the discontinuities.

This discontinuous behavior of $p(N, a, x)$ with respect to x makes it difficult to automate the calculation of the efficacy thresholds $x_0^+(N, a, p_0)$. Therefore for the purposes of this work we simply scanned ranges of x in increments of 0.1% or 0.01% and located the thresholds manually. The corresponding calculations are included in the supplementary material document [22].

2.3. Relationship with the Binomial Proportion Confidence Interval Problem

The last mathematical remark that we wish to comment on is the question of whether anyone has ever proposed or studied an approach like what we have proposed in Section 2. From a mathematical point of view, the answer is yes. Hypothesis testing in which a treatment group is compared against a specific expected adverse outcome probability is the "inverse" (more precisely, the *contrapositive*) of the binomial proportion confidence interval problem, as was noted by Reiczigel [64]. Suppose you run a binomial trial (e.g. tossing several times a possibly loaded coin with a binary success/fail outcome for each toss) and for N trials you observe a failures and $N - a$ successes. Assuming that we know a priori that the failure probability is the same for each trial, what is the unknown probability x of failure? On its face, you could guess that $x = a/N$, but since we are given only a finite sample to work with, the best we can do is to claim that the correct value of x is near the estimate a/N. The challenge of the binomial proportion confidence interval problem is to identify an interval (x_1, x_2) such that we can assert with 95% confidence that $x \in (x_1, x_2)$. A solution to the binomial proportion confidence interval problem can be represented by an incidence function $I(N, a, x, p_0)$ such that $I(N, a, x, p_0) = 1$ if and only if x is inside the interval (x_1, x_2) corresponding to $1 - p_0$ confidence based on the binomial trial sample (N, a). Otherwise, we set $I(N, a, x, p_0) = 0$.

Now, suppose you have a treatment group with N patients and a adverse outcomes and suppose that x is the probability of adverse outcome without treatment. Let $I(N, a, x, p_0)$ represent the solution to the corresponding binomial trial confidence interval problem for a binomial trial with an observed sample (N, a). Then the null hypothesis H_0 implies that the probability x of an adverse outcome without treatment is equal to the probability of adverse outcome with treatment. In turn, that implies $x \in (x_1, x_2)$ with $1 - p_0$ confidence. We obtained thus the Boolean statement

$$H_0 \implies I(N, a, x, p_0) = 1. \tag{10}$$

Using elementary concepts of Boolean algebra, this statement is equivalent to the corresponding contrapositive statement

$$I(N, a, x, p_0) = 0 \implies \neg H_0, \tag{11}$$

Here, $\neg H_0$ represents the logical negation of the null hypothesis H_0. This is the statement of a sufficient condition for rejecting the null hypothesis, and it establishes how the two problems are related to each other. More specifically, what we have previously defined as an efficacy threshold $x_0^+(N, a, p_0)$ is the upper endpoint x_2 of the binomial proportion confidence interval (x_1, x_2). Likewise, the lower endpoint x_1 coincides with the opposite threshold $x_0^-(N, a, p_0)$.

As a result, previous research on the binomial proportion confidence interval problem becomes relevant to our hypothesis testing problem, and in connection to that, we have the following additional remarks. First, the proposed solution $I(N, a, x, p_0)$ should ideally be self-consistent in the sense that the coverage probability $c(N, p_0|x)$ given by

$$c(N, p_0|x) = \sum_{n=0}^{N} I(N, n, x, p_0)\ \mathrm{pr}(N, n|x), \tag{12}$$

should satisfy $c(N, p_0|x) = 1 - p_0$ for all $x \in [0, 1]$. Given a binomial trial with N attempts, the coverage probability $c(N, p_0|x)$, by definition, is the conditional probability that we see any binomial trial outcome (N, n) for which the solution $I(N, n, x, p_0)$ to the corresponding binomial proportion confidence interval problem gives a confidence interval that includes the true binomial trial probability, under the condition that this true probability is equal to x. Second, it can be shown [65] that it is impossible to formulate any such solution $I(N, n, x, p_0)$ that will give the correct coverage probability $c(N, p_0|x)$ for all $x \in [0, 1]$. Therefore, considering that for our purposes it is preferred to overestimate rather than underestimate the efficacy thresholds $x_0^+(N, a, p_0)$, we are happy to settle for a solution that gives conservative coverage such that $c(N, p_0|x) > 1 - p_0$, for all $x \in [0, 1]$.Third, we discover that the overwhelming majority of known methods for solving the binomial proportion confidence interval problem do not have conservative coverage [66]. A notable exception is the Clopper-Pearson interval [67], which gives a corresponding efficacy threshold that reads

$$x_{\mathrm{CP}}(N, a, p_0) = \{x \in [a, N] \mid p_{\mathrm{CP}}(N, a, x) \geq p_0\}, \tag{13}$$

with $p_{\mathrm{CP}}(N, a, x)$ given by Eq. (7). Fig. 3 shows the coverage probability for the Clopper-Pearson interval for sample sizes

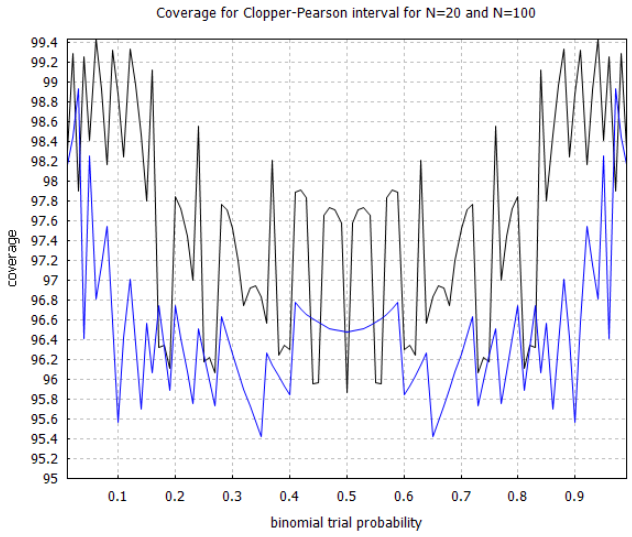

FIG. 3: Coverage probability for the Clopper-Pearson interval [67] with sample sizes $N = 20$ and $N = 100$. The black curve corresponds to $N = 20$ and the blue curve, which is situated below the black curve, corresponds to $N = 100$. The coverage probabilities were calculated using 0.01 increments.

$N = 20$ and $N = 100$. We see that the coverage is indeed conservative and tends to approach 95% from above with increasing sample size.

Although the Clopper-Pearson interval is very well-known and tends to be the go-to method for the binomial proportion confidence interval problem, from the standpoint of hypothesis testing, it cannot be interpreted as a consequence of the two-tail exact Fisher test, under the limit of an infinite control group with a fixed adverse event probability. If we wish to define a solution to the binomial proportion confidence interval problem that corresponds to the aforementioned limit of an exact Fisher test, then we should define

$$I(N, a, x, p_0) = 1 \iff \sum_{n=0}^{N} \mathrm{pr}(N, n|x) H(\mathrm{pr}(N, a|x) - \mathrm{pr}(N, n|x)) \geq p_0. \tag{14}$$

The confidence interval that corresponds to this equation was originally proposed by Sterne [68] and its importance was highlighted more recently by Reiczigel [64].

In Fig. 4, we show the coverage probability for the Sterne interval for sample sizes $N = 20$ and $N = 100$ and note that it also has conservative coverage, which is very desirable in the context of hypothesis testing. In Fig. 5, we compare the coverage probability of the Clopper-Pearson interval against the coverage probability of the Sterne interval and note that although they are both conservative, the Sterne interval has less conservative coverage probability than the Clopper-Pearson, over the same sample size.This suggests that the Sterne interval is the better choice, both from the standpoint of coverage probability and also due to its relationship with the exact Fisher test. On the other hand, due to the zigzag graph of

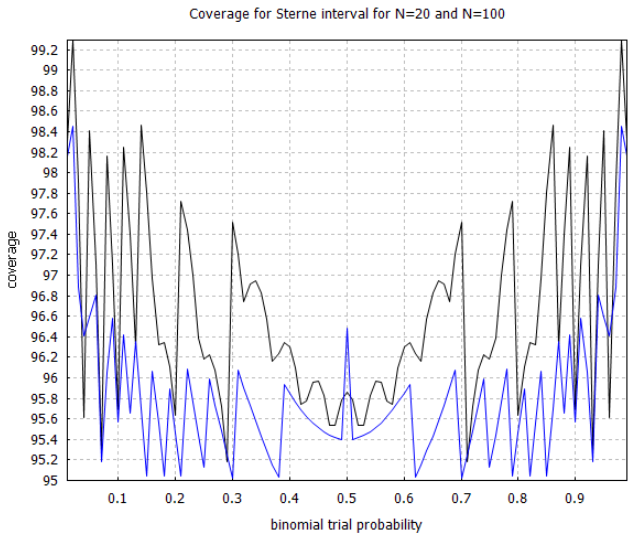

FIG. 4: Coverage probability for the Sterne interval [68] with sample sizes $N = 20$ and $N = 100$. The black curve corresponds to $N = 20$ and the blue curve, which is situated below the black curve, corresponds to $N = 100$. The coverage probabilities were calculated using 0.01 increments.

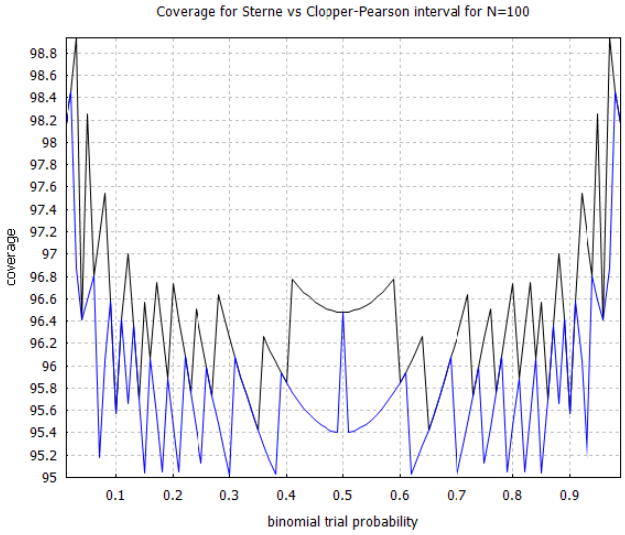

FIG. 5: Comparison of the coverage probability for the Clopper-Pearson interval [67] versus the Sterne interval [68] with sample size $N = 100$. The black curve shows the coverage probability for the Clopper Pearson interval, and the blue curve, which is situated below the black curve, shows the coverage probability for the Sterne interval. The coverage probabilities were calculated using 0.01 increments.

$p(N, a|x)$ as a function of x, a defect of the Sterne interval is that it is not always an interval, but may be punctuated with holes, so there has been some early interest in correcting this problem [69, 70]. Since the Sterne interval has conservative coverage while punctuated with holes, it will have more conservative coverage if one plugs the holes, so there is no need to worry about underestimating the efficacy thresholds.

3. BAYESIAN FACTOR ANALYSIS OF EFFICACY THRESHOLDS

The methodology that we proposed in Section 2 is vulnerable to the criticism that rejecting the null hypothesis, solely on the basis that the p-value satisfies $p < 0.05$, is not sufficient for asserting that treatment efficacy is statistically significant. This is indeed the position of the recent American Statistical Association statement on statistical significance and p-values [6]. The problem is that p-values only measure how incompatible the data are with the null hypothesis. However, this measure does not always do a good job at controlling the probability of a false positive result [71]. To estimate the latter probability we would have to formulate the appropriate alternate hypothesis and consider how much the data is compatible or incompatible with that alternate hypothesis. This has prompted recommendations to lower the p-value threshold down to 0.01 or 0.001 [71, 72]. However, this is only a stopgap measure that does not fundamentally address the problem.

In this section, we supplement the p-value based analysis of Section 2, with a proposal for a Bayesian factor analysis [7–11]. The Bayesian factor compares the alternate hypothesis (treatment efficacy), against the null hypothesis and can be used to calculate the probability of a false positive result [12]. We do not mean to suggest that the Bayesian factor should replace the p-value in hypothesis testing. Our view is that we need to use both. That is, use the p-value to reject the null hypothesis, and then use the Bayesian factor to assess the strength of the evidence in favor of the alternate hypothesis. This viewpoint is similar to earlier proposals for conditional frequentist testing [7]. In the following, we will briefly review the Bayesian factor framework and then outline our specific proposal for validating and adjusting, as needed, the efficacy threshold $x_0^+(N, a, p_0)$.

3.1. Bayesian factor and the false positive rate

Let ,A, B, be two arbitrary events in some probability space. From the definition of conditional probability, we obtain the Bayes rule, noting that

$$p(B|A) = \frac{p(B \cap A)}{p(A)} = \frac{p(A \cap B)}{p(A)} \tag{15}$$

$$= \frac{p(A|B)p(B)}{p(A)}. \tag{16}$$

Let D represent our data, H_0 represent the null hypothesis, and H_1 represent the alternate hypothesis. In the Bayesian statistics framework, we assign probabilities $p(H_0), p(H_1)$ to the hypotheses H_0, H_1 representing our prior belief about how likely each hypothesis is, and then calculate the updated probabilities $p(H_0|D)$ and $p(H_1|D)$ on the condition of observing

the data D. In this way, Bayesian statistics is distinct from frequentist statistics where probabilities are not assigned to the hypotheses themselves.

From the Bayes rule we have,

$$p(H_1|D) = \frac{p(D|H_1)p(H_1)}{p(D)}, \tag{17}$$

$$p(H_0|D) = \frac{p(D|H_0)p(H_0)}{p(D)}, \tag{18}$$

and dividing the two equations gives

$$\frac{p(H_1|D)}{p(H_0|D)} = \frac{p(D|H_1)}{p(D|H_0)}\frac{p(H_1)}{p(H_0)}. \tag{19}$$

The Bayes factor $B(D|H_1, H_0)$ is defined to read

$$B(D|H_1, H_0) = \frac{p(D|H_1)}{p(D|H_0)}, \tag{20}$$

and it is the numerical factor that amplifies our prior belief about the odds ratio $b(H_1, H_0) = p(H_1)/p(H_0)$ after seeing the data D. Here, $p(D|H_1)$ is the probability of seeing the data D if H_1 is true and $p(D|H_0)$ is likewise the probability of seeing the data D if H_0 is true.

To interpret the meaning of the Bayesian factor, the following argument is used to calculate the posterior probabilities $p(H_1|D)$ and $p(H_0|D)$ in terms of $B(D|H_1, H_0)$ and $b(H_1, H_0) = p(H_1)/p(H_0)$. We assume that H_0, H_1 satisfy $p(H_0) + p(H_1) = 1$ and $p(H_0|D) + p(H_1|D) = 1$. Combining the second equation with Eq. (17) and Eq. (18) gives the Bayes theorem

$$p(D) = p(D|H_0)p(H_0) + p(D|H_1)p(H_1), \tag{21}$$

and it follows that the probability of a false positive result is given by

$$p(H_0|D) = \frac{p(D|H_0)p(H_0)}{p(D)} \tag{22}$$

$$= \frac{p(D|H_0)p(H_0)}{p(D|H_0)p(H_0) + p(D|H_1)p(H_1)} \tag{23}$$

$$= \frac{p(D|H_0)p(H_0)}{p(D|H_0)p(H_0)[1 + B(D|H_1, H_0)b(H_1, H_0)]} \tag{24}$$

$$= \frac{1}{1 + B(D|H_1, H_0)b(H_1, H_0)}. \tag{25}$$

We see that the false positive probability approximately scales as the inverse of the Bayes factor $B(D|H_1, H_0)$. On the other hand, the dependence of $p(H_0|D))$ on the prior likelihood ratio $b(H_1, H_0)$, which measures our subjective belief about the odds ratio between H_1 and H_0, before seeing the data D, is uncomfortable. There are three ways to cope with that: First, one can simply join the frequentist camp, consider probabilities based on beliefs as meaningless, and forget about the whole thing. Second, one can use an uninformed prior, meaning that we assume that both hypotheses H_0 and H_1 are equally probable, not having any prior knowledge that favors one over the other, and choose $p(H_0) = p(H_1) = 1/2$, which corresponds to $b(H_1, H_0) = 1$. An interesting third way is to use the reverse Bayesian analysis technique proposed by Colquhoun [12], which is based on the equivalence

$$p(H_0|D) < p_0 \iff b(H_1, H_0) > \frac{1 - p_0}{p_0 B(D|H_1, H_0)}, \tag{26}$$

which relates an upper bound p_0 on the probability $p(H_0|D)$ with a corresponding lower bound $b_{\min}(p_0, B)$ on the prior likelihood ratio $b(H_1, H_0)$, which is given by

$$b_{\min}(p_0, B) = \frac{1 - p_0}{p_0 B}, \tag{27}$$

with B being the value of the corresponding Bayesian factor. The meaning of Eq. (27) is that, given a desired lower bound p_0 for the false positive rate and a threshold B for the Bayesian coefficient, $b_{\min}(p_0, B)$ is the minimum prior likelihood ratio $p(H_1)/p(H_0)$ for our prior knowledge of the extent to which the alternate hypothesis H_1 is favored over the null hypothesis H_0, for which the Bayesian threshold B can control the false positive rate and keep it below p_0. As such, given our subjective choice for $b_{\min}$, one can calculate the threshold B for the Bayesian factor corresponding to the minimum tolerated false positive rate p_0.

Since we wish to constrain the false positive rates to less than 0.05, in order to claim 95% statistical significance, we choose $p_0 = 0.05$. Kass and Raftery [11] and Jeffries [73] both recommend that the threshold $B > 100$ be used for a *decisive* acceptance of the alternate hypothesis H_1 over the null hypothesis H_0. Using $B = 100$ we find that $b_{\min}(0.05, 100) = 0.19$. This means that even if our prior belief is as bad as 5 to 1 in favor of the null hypothesis, a Bayesian factor $B > 100$ is good enough to accept the alternate hypothesis with more than 95% confidence. In this sense, we can indeed claim that $B > 100$ is a reasonable threshold for a *decisive* Bayesian factor. We also know that for $B = (1 - p_0)/p_0$ we have $b_{\min}(p_0, B) = 1$ which corresponds to an uninformed prior likelihood ratio, and for $p_0 = 0.05$, this corresponds to the threshold $B > 19$, which reads $\log B > 1.27$, on a decimal logarithmic scale. Rounding up a bit, we can then argue that $\log B > 1.3$ gives 95% statistical significance for accepting the alternate hypothesis and is a reasonable choice for a threshold for *strong* evidence. As such this threshold is higher than the threshold for strong evidence that was previously proposed by Kass and Raftery [11], and lower than the corresponding threshold, previously proposed by Jeffries [73]. As a practical matter, we shall prefer to use the decisive threshold $\log B > 2$, but for cases that may fall short of that, it is good to know that $1.3 < \log B < 2$ still represents a strong signal.

3.2. Application to hypothesis testing for case series

Now, let us consider how Bayesian factor analysis can be applied to a case series with a treatment group of N patients, where a patients have an adverse outcome. Let x_0 be the corresponding efficacy threshold, determined via the techniques

of Section 2, and let x be the probability of an adverse outcome *with* treatment. We define a null hypothesis H_0 and an alternate hypothesis H_1 about the value of x such that

$$H_0 : x_0 < x \leq 1, \tag{28}$$

$$H_1 : 0 < x \leq x_0. \tag{29}$$

We use for x_0 the upper endpoint of the binomial proportion confidence interval corresponding to the observed data (N, a). Consequently, the null hypothesis H_0 has been defined to place x outside and above that interval, and the alternative hypothesis H_1 considers the remaining possible values for x.

Because both H_0 and H_1 are composite hypotheses, it is necessary to introduce prior probabilities $\mathrm{pr}(x|H_0)$ and $\mathrm{pr}(x|H_1)$, corresponding to H_0 and H_1. It may seem tempting to just use uninformed priors, both for H_0 and H_1, however doing so would certainly not be appropriate for the null hypothesis H_0 in almost all situations, since with many illnesses, we can rule out probabilities of adverse outcome beyond some upper bound p_2. We can thus use instead an uninformed prior on the interval $[x_0, p_2]$, given by

$$\mathrm{pr}(x|H_0(x_0, p_2)) = \begin{cases} 1/(p_2 - x_0), & \text{if } x \in [x_0, p_2] \\ 0, & \text{if } x \in (p_2, 1], \end{cases} \tag{30}$$

and perform an appropriate sensitivity analysis on the parameter p_2. In general, increasing p_2 will tend to increase the Bayes factor, since doing so will tend to increase the contrast between the null and alternate hypotheses. So we can explore how much p_2 can be decreased and still maintain a decisive Bayes factor. Likewise, for the alternate hypothesis H_1, we will use an uninformed prior on the interval $[0, t]$ with $t \leq x_0$ given by

$$\mathrm{pr}(x|H_1(x_0, t)) = \begin{cases} 1/t, & \text{if } x \in [0, t] \\ 0, & \text{if } x \in (t, x_0]. \end{cases} \tag{31}$$

The reason for this choice is that we have found empirically that in some cases, the Bayes factor may actually increase, if instead of an uninformed prior on $[0, x_0]$ we use an uninformed prior on the shorter interval $[0, t]$. From an intuitive standpoint, we surmise that if the data has a very strong efficacy signal, then the contrast between the null and alternate hypotheses is increased when one eliminates the relatively unlikely values of x between t and x_0. For this reason, we shall use the maximum value of the Bayes factor taken over all values $t \in (0, x_0)$, on a decimal logarithmic scale, which is given by

$$b(x_0, p_2) = \max_{t \in (0, x_0]} b_0(x_0, p_2, t), \tag{32}$$

$$b_0(x_0, p_2, t) = \log B(N, a|H_1(x_0, t), H_0(x_0, p_2)). \tag{33}$$

In appendix B we prove that the function $b_0(x_0, p_2, t)$ is initially increasing and then decreasing with respect to t with a maximum in the interval $[a/N, 1]$. If this maximum is located in the narrower interval $[a/N, x_0]$ then the optimal Bayes factor is indeed obtained when we use a choice $t \in (0, p_0)$ for the prior distribution of the alternate hypothesis H_1. If the maximum is formally located at $t > x_0$, then the optimal Bayes factor is obtained at $t = x_0$. The resulting metric $b(x_0, p_2)$ is still dependent on the parameter p_2 of the prior distribution of the null hypothesis H_0.

To complete the metric definition by Eq. (32) and Eq. (33), we now show the calculation of the Bayes factor $B(N, a|H_1(x_0, t), H_0(x_0, p_2))$ between H_1 and H_0 as of function of x_0, p_2, t and the data N, a. We note that the probabilities for seeing the data (N, a) under the hypotheses H_1 and H_0 are given by:

$$\mathrm{pr}(N, a|H_0(x_0, p_2)) = \int_{x_0}^{1} \mathrm{d}x \ \mathrm{pr}(N, a|x) \ \mathrm{pr}(x|H_0(x_0, p_2)) \tag{34}$$

$$= \frac{1}{p_2 - x_0} \int_{x_0}^{p_2} \mathrm{d}x \ \mathrm{pr}(N, a|x) \tag{35}$$

$$= \frac{1}{p_2 - x_0} \binom{N}{a} \int_{x_0}^{p_2} x^a (1 - x)^{N-a} \ \mathrm{d}x, \tag{36}$$

and

$$\mathrm{pr}(N, a|H_1(x_0, p_2)) = \int_{0}^{x_0} \mathrm{d}x \ \mathrm{pr}(N, a|x) \ \mathrm{pr}(x|H_1(x_0, t)) \tag{37}$$

$$= \frac{1}{t} \int_{0}^{t} \mathrm{d}x \ \mathrm{pr}(N, a|x) \tag{38}$$

$$= \frac{1}{t} \binom{N}{a} \int_{0}^{t} x^a (1 - x)^{N-a} \ \mathrm{d}x, \tag{39}$$

consequently, the corresponding Bayes factor is given by

$$B(N, a|H_1(x_0, t), H_0(x_0, p_2)) = \frac{\mathrm{pr}(N, a|H_1(x_0, p_2))}{\mathrm{pr}(N, a|H_0(x_0, p_2))} \tag{40}$$

$$= \frac{p_2 - x_0}{t} \frac{\displaystyle\int_0^t x^a (1 - x)^{N-a} \ \mathrm{d}x}{\displaystyle\int_{x_0}^{p_2} x^a (1 - x)^{N-a} \ \mathrm{d}x}. \tag{41}$$

The integrals can be calculated using exact algebra or numerically with the open source computer algebra software Maxima [74]. The exact algebra calculation takes longer to carry out, but we have confirmed that the numerical calculation using the function quad_qagr is just as accurate.

In order to control for the false positive rate, we propose that the efficacy thresholds $x_0^+(N, a, p_0)$ with $p_0 = 0.05$ should be increased, if necessary, by requiring that they also satisfy $b(x_0, p_2) \geq 2$. Since the threshold used for a decisive Bayes factor with $p_0 = 0.05$ corresponds approximately to $b_{\min}(p_0, B) = 1/5$, it is reasonable to use the empirical formula

$$b(x_0, p_2) \geq \log\left(\frac{5(1 - p_0)}{p_0}\right), \tag{42}$$

to adjust the efficacy thresholds $x_0^+(N, a, p_0)$ for an arbitrary value of demanded confidence p_0. For $p_0 = 0.01$, this gives $b(x_0, p_2) \geq 2.7$ and for $p_0 = 0.001$ we find $b(x_0, p_2) \geq 3.7$ as the Bayes factor thresholds corresponding to a prior likelihood ratio $p(H_1)/p(H_0) = 1/5$ and, as such, they are the

thresholds that we recommend imposing on the Bayes factors for the purpose of adjusting the corresponding efficacy thresholds $x_0^+(N, a, p_0)$ for the choices $p_0 = 0.01$ and $p_0 = 0.001$.

4. APPLICATION TO THE ANALYSIS OF EARLY OUTPATIENT COVID-19 TREATMENT CASE SERIES

4.1. Review of the Zelenko, Procter and Raoult case series

We shall now analyze, using the aforementioned methodology, the high-risk patient case series by Zelenko [2, 13, 14], Procter [15, 16], and Raoult [17]. The main reason for focusing on these case series specifically, is that they consist exclusively of high-risk patients, where early outpatient treatment is expected to make a difference. In this subsection we shall present the details of these case series, focusing on outcomes and treatment protocols used.

In the Zelenko April 2020 letter [13], Zelenko reported on his outcomes based on a total of 1,450 patients that he treated for COVID-19 until April 28, 2020 in an Orthodox Jewish community in upstate New York. From this cohort, 405 patients were classified as high risk and treated with his triple-drug therapy (hydroxychloroquine, azithromycin, zinc sulfate). The reported outcomes were 6 hospitalizations and 2 deaths. From amongst the patients classified as low risk, who were given only supportive care, there were no hospitalizations or deaths. Zelenko's criteria for risk stratification define three categories of high risk patients: (1) every patient older than 60; (2) every patient younger than 60 but with comorbidities; (3) patients younger than 60 and without comorbidities that presented with shortness of breath.

A subset of the April 28, 2020 case series was published in a case controlled study [2] that included only the treated patients with COVID-19 infection that was confirmed by a PCR test or an antibody IgG test. The remaining patients were clinically diagnosed from symptomatic presentation and via ruling out a bacterial or influenza infection. This Derwand–Scholz–Zelenko study (hereafter DSZ study) [2] included 335 patients of which, 141 patients were classified as high-risk patients and treated with the triple drug protocol with 4 hospitalizations and 1 death. Detailed demographic data is given for the high-risk patient treatment group, including a detailed breakdown in the three high-risk categories. The study also included a control group of 377 patients who were seen by other treatment centers in the same community, that were only offered supportive care and no early outpatient treatment. From this untreated group, 13 patients died and 58 patients were hospitalized. The untreated group includes both low-risk and high-risk patients, so we expect that it underestimates both the hospitalization and mortality risk for high-risk patients. Unfortunately, demographic data was not available for the untreated group, so from a strictly methodological point of view, one cannot entirely rule out the theoretical possibility that the untreated group might have consisted of patients that are at higher risk on average than those of the high-risk treatment group. On the other hand, using a case series of untreated patients from Israel [55], with demographic data indicating a combination of low and high-risk patients, with 143 deaths reported out of 4,179 untreated patients, gives the same mortality rate as in the DSZ control group, suggesting that the DSZ control group also consists of a mixed demographic of low and high risk patients.

The June 2020 Zelenko case series [14] is reported in a letter that Zelenko sent to the Israeli Health Minister at the time, Dr. Moshe Bar Siman-Tov, on June 14, 2020, which was later made publicly available. In the letter, Zelenko reported that a total of approximately 2,200 patients were seen as of June 14, 2020, with 800 patients deemed high-risk under the same criteria and treated with the triple-drug therapy, since the beginning of the pandemic. The reported cumulative outcomes are, 12 hospitalizations, 2 deaths, no serious side effects, and no cardiac arrhythmias.

During the April 2020–June 2020 interval, Zelenko enhanced his triple drug therapy protocol with oral dexamethasone and budesonide nebulizer at the beginning of May 2020. He introduced the blood thinner Eliquis towards the end of May 2020 and beginning of June 2020. Ivermectin was not used by Zelenko until October 2020. Consequently, the DSZ study [2] and the Zelenko April 2020 case series [13] reflect the outcomes of the triple drug therapy, when used by itself as an early outpatient treatment. The Zelenko June 2020 case series [14] includes the use of steroid medications and a blood thinner, so the underlying treatment protocol is closer to the McCullough protocol [18–20].

It is worth noting that both letters [13, 14] were originally posted on Google Drive by Zelenko and were censored by Google during 2021. The April 2020 letter [13] was cited by Risch [59], whose paper has also preserved the corresponding case series data. The June 2020 letter case series data [14] was independently reported by a subsequent publication by Risch [75], however, it included only the number of reported deaths, and not the number of hospitalizations. The authors have attached copies of all three Zelenko letters [13, 14, 21] to our supplementary material document [22].

The Procter case series were reported consecutively in two publications [15, 16]. The first paper [15] reports on 922 patients that were seen between April 2020 and September 2020, of which 320 were risk stratified as high-risk patients and treated with the McCullough protocol [18–20]. The outcome was, 6 hospitalizations and 1 death. The second paper [16] reports on an additional patient cohort seen between September 2020 and December 2020. Out of the total number of patients during that time period, 549 were risk stratified as high-risk and treated with an outcome of 14 hospitalizations and one death. For both case series, the risk stratification criteria were similar to those used by Zelenko. However, the age threshold used to risk stratify patients as high-risk was lowered to 50 years. The medications used were customized for each patient in accordance with the McCullough protocol [18–20] and included hydroxychloroquine, ivermectin, zinc, azithromycin, doxycycline, budesonide, foliate, thiamin, IV fluids, and for more severe cases, dexamethasone and ceftriaxone were also added. Demographic details for the cohorts were reported in the respective publications [15, 16].

The final high-risk patient case series is extracted from a

Study	Total	High-risk	Hospitalizations & Deaths	
DSZ study [2]	712	141	4 (2.8%)	1 (0.7%)
Zelenko April 2020 [13]	1450	405	6 (1.4%)	2 (0.4%)
Zelenko June 2020 [14]	2200	800	12 (1.5%)	2 (0.25%)
Procter I [15]	922	320	6 (1.8%)	1 (0.3%)
Procter II [16]	?	869	20 (2.3%)	2 (0.2%)
Raoult [17]	10429	1495	106 (7.0%)	5 (0.3%)
DSZ control [2]	377	< 377	58 (>15%)	13 (>3.4%)
Israeli control [55]	4179	< 4179	N/A	143 (>3.4%)
Raoult control [17]	2114	520	38 (7.3%)	11 (2%)

FIG. 6: Case series list: The table lists the total number of patients, the subset of high risk patients that were treated with a sequenced multidrug regimen, number of patients that were hospitalized, and number of deaths, for the following case series: Derwand-Scholtz-Zelenko study treatment group [2], Zelenko's complete April 2020 data set [13], Zelenko's complete June 2020 data set [14], Procter's observational studies [15, 16], and Raoult's high risk (older than 60) treatment group [17]. The table also lists the same data for the control group in the DSZ study [2], the untreated group in the Israeli study [55], and the control group in the Raoult study [17].

Study	odds ratio	95% CI	p-value
Exact Fisher tests on mortality rates			
DSZ study vs DSZ control	0.2	0.02–1.54	**0.12**
Zelenko April 2020 vs DSZ control	0.13	0.03–0.61	0.003
Zelenko June 2020 vs DSZ control	0.07	0.01–0.31	10^{-5}
DSZ vs Israeli control	0.2	0.03–1.45	**0.09**
Zelenko April 2020 vs Israeli control	0.14	0.03–0.57	0.0002
Zelenko June 2020 vs Israeli control	0.07	0.02–0.28	10^{-9}
Exact Fisher tests on hospitalization rates			
DSZ vs DSZ control	0.16	0.05–0.45	0.02
Zelenko April 2020 vs DSZ control	0.08	0.03–0.19	10^{-13}
Zelenko June 2020 vs DSZ control	0.08	0.04–0.16	10^{-19}

FIG. 7: Exact Fisher test comparing the mortality rate reduction and hospitalization rate reduction between the high risk patient treated group the DSZ study [2], Zelenko's complete April 2020 data set [13], and Zelenko's complete June 2020 data set [14] against the low risk and high risk patient control groups in the DSZ study [2] and the Israeli study [55]. The p-values where there is a failure to establish 95% confidence are highlighted.

recent cohort study [17] of 10,429 patients that were seen between March 2020 and December 2020 by Raoult's IHU Méditerranée Infection hospital in Marseille, France. From the entire cohort, 8,315 patients were treated with hydroxychloroquine, azithromycin and zinc. Of those patients, those older than 70 or with comorbidities were also treated with enoxaparin and low-dose dexamethasone was given on a case by case basis to patients that presented with inflammatory pneumonopathy, high viral loads, or on a case by case basis. This treatment protocol is consistent to some extent with the principles that underlie the McCullough protocol [18–20]. The remaining 2,114 patients did not receive hydroxychloroquine or azithromycin or both because it was either contraindicated or because the patients did not consent to using one or two of these medications. This cohort was used in the Raoult study [17] as a control group. The study risk-stratified the patients by age (see Table 1 of Ref. [17]) making it possible to extract a case series of high-risk patients under the restriction age $\geq$ 60. In the treatment group, this results in 1,495 high-risk patients with 5 deaths and 106 hospitalizations. In the control group, under the age $\geq$ 60 constraint, there are 520 high-risk patients with 38 hospitalizations and 11 deaths. The authors note that no serious adverse events to the medications were reported and that the reported deaths were not related to side effects of hydroxychloroquine or azithromycin. Furthermore, no deaths were reported for age < 60 cohort in both the treatment group and control group.

Fig. 6 summarizes the aforementioned case series, including the treatment groups from the DSZ studu [2], the Zelenko [2, 13, 14] and Procter [15, 16] case series and the age $\geq$ 60 treatment group from the Raoult study [17]. Note that the Zelenko June 2020 case series and the Procter II case series as reported on Fig. 6, combine the two respective consecutive case series. We also report on Fig. 6 the DSZ study's control group [2], the alternative Israeli control group [55] and the age $\geq$ 60 part of the Raoult control group [17]. We emphasize that all reported treatment group case series consist of high-risk patients.

From a cursory examination of Fig. 6, we see that the mortality rate is consistent across all treatment groups, which speaks to the consistency Bradford Hill criterion [76]. Hospitalization rates are also consistent between the Zelenko [2, 13, 14] and Procter case series [15, 16], but there is a clear discrepancy with the hospitalization rates reported in the Raoult treatment case series [17]. We believe that the reason for the discrepancy is that both Zelenko and Procter explicitly aimed to prevent hospitalizations due to the poor outcomes of the inpatient treatment protocols used in the United States. In Marseille, France, Raoult had the option of using his IHU Méditerranée Infection hospital, for short hospitalizations, in order to closely monitor his more concerning cases.

In Fig. 7, we show the results of comparing the Zelenko April 2020 [13] and Zelenko June 2020 [14] case series against both the original DSZ control group [2] as well as the alternative control group from Israel [55]. Although in the original DSZ study [2] mortality rate reduction was not statistically significant, we have found that comparing either the Zelenko April 2020 case series [13] or the June 2020 case series [14] against either control group, gives more than 90% mortality rate reduction, which is also statistically significant in terms of both p-value and confidence interval. Likewise, we see at least 90% hospitalization rate reduction when the Zelenko April 2020 case series or Zelenko June 2020 case series is compared against the DSZ control group, which is statistically significant as well. Because the control groups consist of a combination of both low-risk and high-risk patients, whereas the treatment groups consist of only high-risk patients, the resulting comparisons are biased towards the null, and are thus

Study	95% threshold	99% threshold	99.9% threshold
Mortality rate efficacy thresholds			
DSZ study	3.8% (3.9%)	5.3% (5.2%)	7.0% (6.9%)
Zelenko April 2020	1.8%	2.4% (2.3%)	2.9%
Zelenko June 2020	1.0%	1.2% (1.2%)	1.6% (1.5%)
Procter I	1.7% (1.8%)	2.3%	3.1%
Procter II	0.84% (0.83%)	1.08% (1.07%)	1.4% (1.38%)
Raoult	0.79% (0.78%)	0.96% (0.95%)	1.18%
Hospitalization rate efficacy thresholds			
DSZ study	7.0% (7.2%)	8.8% (8.7%)	10.6% (10.7%)
Zelenko April 2020	3.2%	3.9%	4.7%
Zelenko June 2020	2.7%	3.0%	3.5%
Procter I	4.1%	4.9%	5.9%
Procter II	3.6%	4.0%	4.5%

FIG. 8: Mortality and hospitalization rate reduction efficacy thresholds, defined as the upper end of the Sterne interval [68], corresponding to 95%, 99%, and 99.9% confidence, for the DSZ study treatment group [2], Zelenko's complete April 2020 data set [13], Zelenko's complete June 2020 data set [14], Procter's observational studies [15, 16], and Raoult's high risk (older than 60) treatment group [17]. The approximate efficacy thresholds obtained by the upper endpoint of the Clopper-Pearson interval [67] are shown in parenthesis when not equal to the Sterne interval [68] threshold.

Age	Deaths	Cases	CFR
10-19	0	416	0%
20-29	7	3619	0.193%
30-39	18	7600	0.237%
40-49	38	8571	0.4%
50-59	130	10008	1.3%
60-69	309	8583	3.6%
70-79	312	3918	7.96%
≥ 80	208	1408	14.8%
≥ 60	829	13909	5.96%

FIG. 9: Crude Case Fatality Rate data, in the absense of early outpatient treatment, based on early data from China as of February 11, 2020, and published on March 30, 2020. [54]

underestimating the actual efficacy of the respective treatment protocols.

4.2. Case series efficacy thresholds

We have calculated the efficacy threshold for mortality rate reduction and hospitalization rate reduction corresponding to the case series by Zelenko [2, 13, 14], Procter [15, 16], and Raoult [17]. The calculations are shown in the supplementary material document [22]. The results are tabulated in Fig. 8. We display the efficacy thresholds for 95%, 99% and 99.9% confidence, which are calculated as the upper end points of the corresponding Sterne interval [68] (see Eq. (4)), and in parenthesis, we display the approximate thresholds, obtained using the upper endpoint of the corresponding Clopper-Pearson interval [67], (see Eq. (13)), when they diverge from the exact thresholds by more than the precision used. We use precision of 0.1% for most case series, except for the two largest ones, Procter II [16] and Raoult [17], where we use 0.01% precision.

Each threshold corresponds to a mathematically rigorous conditional statement about rejecting the null hypothesis that the corresponding early outpatient treatment protocol is ineffective. For example, the 1.7% efficacy threshold corresponding to 95% confidence for rejecting the null hypothesis in the Zelenko April 2020 case series corresponds to the following statement: *if the expected mortality rate for an equivalent cohort without early outpatient treatment exceeds 1.7%, then the null hypothesis can be rejected with at least 95% confidence*. Similar statements can be formulated for each efficacy threshold metric on Fig. 8. These statements are mathematical facts. However, to complete the inference argument, they need to be paired with an inevitably subjective statement that provides an estimate, or at least a lower bound, on the expected mortality or hospitalization rates of similar cohorts without early outpatient treatment. If we can assert that these rates are in an interval that is entirely above the corresponding efficacy thresholds, then we can reject the null hypothesis. Secondarily, we need an inference about the intervals of mortality or hospitalization rates, in the absence of early outpatient treatment, in order to do the Bayesian adjustment of the efficacy thresholds.

In general, patients have been classified as high-risk based on the following three categories: (1) old age; (2) comorbidities or obesity (with BMI $\geq 30\text{kg/m}^2$); (3) shortness of breath upon presentation. The age threshold for high risk classification is age ≥ 60 for the Zelenko [2, 13, 14] and Raoult [17] case series, and age ≥ 50 for the Procter [15, 16] case series. The high-risk treatment groups for the Zelenko [2, 13, 14] and Procter [15, 16] case series include the demographic distribution of all three categories of high-risk patients, whereas in the Raoult [17] case series we have included only age ≥ 60 patients. Our approach, in the following, is to lower bound the mortality rate, in the absence of early outpatient treatment, separately for each of the three high-risk patient categories. Then, the common lower bound becomes applicable to any demographic distribution of the three categories. To establish the existence of treatment efficacy, it is sufficient for this lower bound to exceed the corresponding efficacy thresholds of Fig. 8. In the following, we shall now consider the mortality rate for each of the three high-risk patient categories separately.

With regards to the first category of patients classified as high-risk due to old age, the earliest data from China [54] as of February 11, 2020, estimated a minimum of 3.6% mortality rate for patients older than 60 and a minimum of 1.3% mortality rate for patients older than 50 (see Fig. 9). These numbers are corroborated with numbers from China [52] and Italy [53] as of March 17, 2020 (see Fig. 10). The 3.6% mor-

Age	Italy CFR	China CFR
0-9	0%	0%
10-19	0%	0.2%
20-29	0%	0.2%
30-39	0.3%	0.2%
40-49	0.4%	0.4%
50-59	1.0%	1.3%
60-69	3.5%	3.6%
70-79	12.8%	8.0%
≥ 80	20.2%	14.8%

FIG. 10: Crude Case Fatality Rate data, in the absense of early outpatient treatment, based on early data from China and Italy as of March 17, 2020 and published on March 23, 2020 [52, 53].

Comorbidity CFR from Chinese study [52]			
Comorbidity	Deaths	Cases	CFR
Cardiovascular disease	92	873	10.5%
Diabetes	80	1102	7.3%
Respiratory disease	32	511	6.3%
Hypertension	161	2683	6%
Cancer	6	107	5.6%
Comorbidity CFR from Israeli study [55]			
Comorbidity	Deaths	Cases	CFR
Cardiovascular disease	87	518	16.7%
Diabetes	71	531	13%
Respiratory disease	23	361	6%
Hypertension	102	744	13.7%
Cancer	37	264	10%

FIG. 11: Case fatality rate based on early-stage analysis of COVID-19 outbreak in China in the period up to February 11, 2020 [52] vs similar statistics from Israel published on September 7, 2020 [55].

tality rate for age ≥ 60 exceeds the 95% efficacy thresholds for the Zelenko April 2020 [13] and Zelenko June 2020 case series [14] by a wide margin. The gap with the efficacy threshold of the DSZ study [2] is too small due to the small sample size of patients. The most noteworthy comparison is with the Raoult case series [17] which consists of exclusively patients with age ≥ 60. There is a large gap between 3.6% and the 95% efficacy threshold, which is down to 0.79%, and even the 99.9% efficacy threshold, which is at 1.16%. So there is an unambiguous, very strong signal of benefit with the Raoult case series [17] with respect to mortality rate reduction. With the Procter I [15] and Procter II [16] case series, the age threshold used for risk stratification was age ≥ 50 and for the Procter II case series [16], there is a small gap between the 0.84% efficacy threshold and a 1.3% lower bound on the untreated mortality rate. A more favorable comparison is possible if one uses the United States case fatality rate (hereafter CFR) [77], or adjusted data from the CDC [56–58] which will be discussed at the end of this subsection.

The second category of high risk patients are patients with comorbidities regardless of age. In Fig. 11, we show case fatality rates with respect to comorbidities (i.e. cardiovascular disease, diabetes, respiratory disease, hypertension, cancer), based on data from China [52] in the period up to February 11, 2020, and additional data from Israel [55] with patients diagnosed in the period up to April 16, 2020, and deaths recorded up to July 16, 2020. There is variability in mortality rates from 5% to 15%, with the entire interval clearly exceeding, by a very wide margin, the efficacy thresholds, for all case series reported on Fig. 8. The Israeli data appear to show higher mortality rates than the data from China, and the reason for that could be that the Israeli study [55] accounted for the time lag between patient diagnosis and death.

These studies do not account for the mortality risk from obesity and also do not account for the mortality risk corresponding to the third category of high-risk patients that present with shortness of breath. A collaborative study by Risch and a research group in Brazil [78], found, using multivariate regression analysis, that both obesity and dyspnea pose a higher mortality risk than heart disease (see Table 2 of Ref. [78]), therefore, we expect that they both lie in the same 5% to 15% interval as patients with other comorbidities.

For the case of obesity, as a mortality risk factor, this conclusion is also supported by more recent meta-analysis [79], showing that obesity is a greater mortality risk factor than diabetes and hypertension, and one that increases with increasing BMI. A study of 148,494 patients across 238 hospitals by the CDC [80] also confirms that obesity is an increasing mortality risk factor with increasing BMI. It is known that obesity is associated with increased levels of the inflammatory cytokines TNF-α (tumor necrosis factor alpha), IL-1β (interleukin-1-beta), and IL-6 (interleukin 6), produced by macrophages in the adipose tissue [81]. A study of 9390 hospitalized patients in Abu Dhabi, United Arab Emirates, has found that patients with severe COVID-19 symptoms, requiring intensive care, had significantly elevated IL-6 biomarker relative to patients that presented with mild or moderate symptoms [82]. An earlier meta-analysis [83] has also confirmed that the IL-6 biomarker is associated with severe progression of the COVID-19 disease. Consequently, there is a very compelling biological mechanism that explains why obesity is a severe risk factor for progression of the disease to the covid-19 pneumonia phase, requiring a high risk classification and immediate early outpatient treatment.

For the case of patients presenting with shortness of breath, it is important to appreciate the fact that, without an early outpatient treatment intervention, such presentation implies that the disease is progressing beyond the viral replication phase, into the COVID-19 pneumonia phase, soon to be followed with the thromboembolic stage, oxygen desaturation, and hospitalization. It is thus self-evident that these patients should be classified as high-risk and treated immediately. Assuming that most of such patients will be hospitalized without outpatient treatment, we can also estimate the corresponding mortality risk, in the absence of outpatient treatment, by looking at

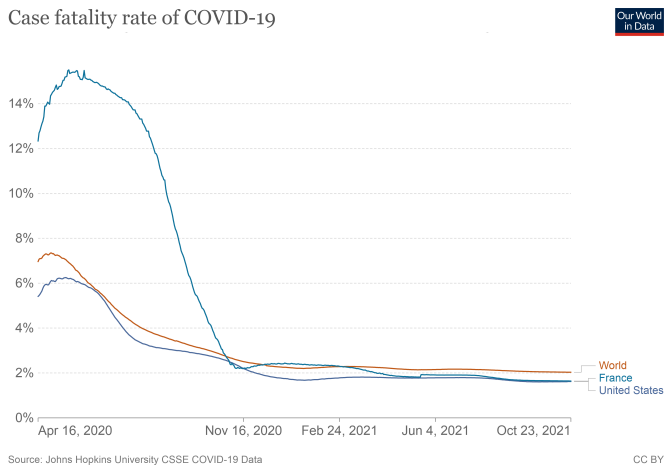

FIG. 12: Cumulative case fatality rate in the United States and France between April 2020 and November 2021.

the conditional probability of death, assuming hospitalization has already taken place. A study by the Houston Methodist Hospital [84] has shown an average mortality rate of 5.8% for hospitalized patients between March 2020 and July 2020, in spite of the use of hydroxychloroquine and anticoagulants. Furthermore, the study reports 12.1% mortality rate, for hospitalized patients between March 13th 2020 and May 15th 2020, and 3.5% mortality rate between May 16th 2020 and July 7, 2020, corresponding to two consecutive surges, noting that the second surge targeted younger patients than the first surge. A prospective multicenter study [85] from Italy of 1050 patients in the Coracle registry, between February 22, 2020 and April 1, 2020, showed an overall 13% average mortality rate, and more specifically, 7.4% mortality rate for hospitalized patients that do not require supplemental oxygen or invasive ventilation, 12.8% mortality rate for hospitalized patients that require supplemental oxygen, and 22.9% mortality rate for hospitalized patients that are invasively ventilated. In light of the above mortality rates for hospitalized patients, the mortality risk of patients presenting with shortness of breath exceeds the efficacy of thresholds of Fig. 8, if we expect that more than half of them will be hospitalized, in the absence of any attempt at early outpatient treatment.

Based on the above arguments, we can lower bound the untreated mortality risk by 3% for each of the three categories of high-risk patients, using age $\geq$ 60 threshold for the first category. Consequently this lower bound can be applied to any demographic distribution between these three categories, for any particular high-risk patient cohort, and it is sufficiently high to exceed the efficacy threshold for the Zelenko April 2020 [13], Zelenko June 2020 [14] and Raoult case series [17].

A completely different approach is to compare the efficacy thresholds against the CFR for the entire population [77]. The CFR for the United States and France is displayed on Fig. 12 for the time period between April 2020 and October 2021. During 2020, the CFR ranged from 2% to 6% in the United States and from 2% to 16% in France. In both countries, the CFR converged to 1.7% during 2021 and remained roughly constant, with very small oscillations throughout 2021. The minimum value of 1.7% exceeds the mortality rate reduction efficacy thresholds for the Zelenko June 2020 [14], Procter II[16], and Raoult case series [17]. Taking the CFR at face value, this is a very strong signal of efficacy, because the CFR includes asymptomatic, low-risk, and high-risk patients, regardless of whether they received early treatment, against solely high-risk patients in the treatment groups of the respective case series. This comparison strongly biases against being able to reject the null hypothesis, and nevertheless, we are still able to do so.

In particular, we note that in the United States, the CFR ranged from 2% to 6% during 2020, which lies above the 1.8% mortality rate reduction efficacy threshold for Zelenko April 2020 case series [13]. This was a strong indicator in favor of adopting Zelenko's triple-drug protocol at that time on an emergency basis, but was unfortunately not implemented in the United States for outpatients [86]. By June 2020, the respective efficacy threshold decreased to 1.0% while the CFR was still in the neighborhood of 3%. Thus, there was a very strong signal in favor of adopting the Zelenko triple-drug therapy by the summer of 2020.

Alternatively to using the CFR, we can also estimate the mortality risk of the first category of high risk patients (age $\geq$ 60 or age $\geq$ 50) using adjusted estimates by the CDC [56–58] of COVID-19 deaths per symptomatic cases. The CDC report attempts to adjust for the differences in underreporting of symptomatic illness, hospitalizations, and deaths, and it is based on reports ranging from February 2020 to September 2021. The raw data and a copy of the CDC Report website are given in our supplementary material document [22]. From that, we calculate for the age $\geq$ 50 group a mortality rate of 2.26% (95% CI: 1.94% – 2.61%) which exceeds by a wide margin the 95% efficacy thresholds for mortality rate reduction for both Procter I [15] and Procter II [16] case series, and for the Zelenko April 2020 [13], Zelenko June 2020 [14] and Raoult case series [17]. We cannot deduce an age $\geq$ 60 mortality rate from the CDC data, but note that the age $\geq$ 65 mortality rate, according to the CDC is 4.79% (95% CI: 4.11% to 5.52%).

4.3. Analysis of hospitalization rate reduction efficacy

In Fig. 8, we see that the 95% efficacy thresholds for hospitalization rate reduction range from 2.7% to 4.1% for all case series, with the exception of the DSZ case series, where it is at 7.0%, due to the smaller sample size. These thresholds can be compared against the following empirical data. At the beginning of the pandemic, based on data from China until February 11, 2020, there was an initial estimate [54] that the probability of hospitalization for a high-risk age $\geq$ 60 cohort would range from 10% to 18%. The control group from Zelenko's study [2] consisting of both low and high-risk patients, again at the beginning of the pandemic here in the United States, reported 377 patients with 58 hospitalizations, corresponding to 15% hospitalization rate. In the Cleveland study [87], which

was used to train a predictive model for the risk of hospitalization and death based on patient medical history, the entire dataset consisted of a total of 4,536 patients between March 8, 2020 and June 5, 2020. There were 582 hospitalizations corresponding to 21% hospitalization rate. In the Mass General Brigham hospital study [88], from a cohort of 12,347 patients that tested positive, there were 3,401 hospitalizations between March 4, 2020 and July 14, 2020, corresponding to a 27% hospitalization rate. This was also a cohort that included both low-risk and high-risk patients. The CDC adjusted data [56–58] between February 2020 and September 2021, estimate 13.79% (95% CI: 17.09% to 28.52%) hospitalization probability for the age $\geq$ 50 group, given a symptomatic infection. For the age $\geq$ 65 cohort, this estimate increases to 22.09% (95% CI: 17.09% to 28.52%)

Overall, our observation is that we tend to see numbers ranging from 10% to 28% with substantial variability between various cohorts, all of which were not given early outpatient treatment. On the other hand, we see that the case series of high risk patients shown in Fig. 8, have efficacy thresholds for hospitalization rate reduction ranging from 2.7% to 4.1%, which have a substantial separation from the 10% to 28% interval. We interpret this big gap between the two intervals as strong evidence of the existence of hospitalization rate reduction efficacy as a result of the respective early outpatient treatment protocols in the Zelenko April 2020 [13], Zelenko June 2020 [14], Procter I [15], Procter II case series [16]

Bayes factors at the mortality rate efficacy thresholds				
Study	95% threshold	$p_2 = 0.02$	$p_2 = 0.05$	$p_2 = 0.1$
DSZ study	3.8%	N/A	1.38	1.99
Zelenko April 2020	1.8%	1.17	2.04	2.45
Zelenko June 2020	1.0%	2.06	2.66	3.02
Procter I	1.7%	1.28	2.07	2.47
Procter II	0.84%	1.92	2.48	2.82
Raoult	0.79%	1.91	2.45	2.79
Bayes factors at the hospitalization rate efficacy thresholds				
Study	95% threshold	$p_2 = 0.02$	$p_2 = 0.05$	$p_2 = 0.1$
DSZ study	7.0%	1.30	1.71	1.92
Zelenko April 2020	3.2%	2.00	2.24	2.39
Zelenko June 2020	2.7%	2.24	2.47	2.61
Procter I	4.1%	1.89	2.15	2.32
Procter II	3.6%	1.98	2.23	2.39

FIG. 13: Bayes factor (decimal logarithm) corresponding to the 95% efficacy threshold (Sterne interval [68]) for mortality and hospitalization rate reduction, using maximum untreated mortality rate p_2 for high risk patients at $p_2 \in \{0.02, 0.05, 0.10\}$ and maximum untreated hospitalization rate p_2 for high risk patients at $p_2 \in \{0.10, 0.15, 0.20\}$, for the DSZ study treatment group [2], Zelenko's complete April 2020 data set [13], Zelenko's complete June 2020 data set [14], Procter's observational studies [15, 16], and Raoult's high risk (older than 60) treatment group [17].

4.4. Bayesian analysis of efficacy thresholds

We shall now assess whether the efficacy thresholds need to be increased, using the Bayesian technique described in Section 3, in order to control the false positive rate. In Fig. 13, we have calculated the logarithmic Bayesian metric $b(x_0, p_2)$, given by Eq. (32), for the mortality and hospitalization rate redaction efficacy thresholds corresponding to 95% confidence, using a range of values of p_2 for the purpose of sensitivity analysis. The calculation details are available in our supplementary material document [22]. Recall from Section 3, that p_2 corresponds to our sense of the worst possible probability of the respective adverse outcome (hospitalization or death) in high-risk patients in the absence of early outpatient treatment. As such, 5% to 10% is a typical range for mortality rates in untreated high-risk patients, making $p_2 = 5\%$ a highly conservative choice. We did not consider values higher than 10%, even though worse probabilities are possible, because for $p_2 > 10\%$, we see that all logarithmic Bayesian factors already satisfy $b(x_0, p_2) \geq 2$. We have also looked at $p_2 = 2\%$, which is obviously entirely unrealistic, because it corresponds to the mortality rate of the Raoult control group [17] where some partial treatment was given.

In spite of that, we see that the logarithmic Bayesian factor $b(x_0, p_2)$ for $p_2 = 2\%$ is very close to 2 for the Procter II [16], and Raoult [17] case series, and for the Zelenko June 2020 case series [14], it is above 2. For the Zelenko April 2020 [13] and Procter I [15] case series, the logarithmic Bayesian factor is too small for $p_2 = 2\%$ to control the false positive rate for mortality rate reduction, however it is above the decisive threshold for $p_2 = 5\%$ and $p_2 = 10\%$. This is not a serious concern, since that particular choice of $p_2 = 2\%$ is unrealistically small. For the DSZ study [2], a $p_2 = 2\%$ logarithmic Bayesian factor is not relevant since the corresponding p-value efficacy threshold exceeds 2%. Furthermore, the $p_2 = 5\%$ logarithmic Bayesian factor is also far below the decisive threshold and the $p_2 = 10\%$ logarithmic Bayesian factor is only borderline decisive. This signals that the sample size in the DSZ study [2] maybe too small to be used to establish a statistically significant mortality rate reduction in terms of controlling the false positive rate, even if we are able to successfully reject the null hypothesis.

Likewise, for the hospitalization rate reduction efficacy thresholds, we have used the values $p_2 = 10\%, 15\%, 20\%$ based on our expectation of a typical 10% to 28% range for the probability of hospitalization, in the absence of early outpatient treatment. We did not consider $p_2 > 20\%$ since almost all of the logarithmic Bayesian factors satisfy $b(x_0, p_2) \geq 2$ at $p_2 = 20\%$. For all case series, except for the DSZ study [2] and the Procter I case series [15], the numbers are good for all three values of p_2, therefore, it is not necessary to increase the 95% confidence efficacy threshold for hospitalization rate reduction. For the DSZ study [2] we see that the logarithmic Bayesian factors stand out as being noticeably below 2. This is caused by the small sample size, and it signals that although the 7.0% threshold is appropriate for rejecting the null hypothesis, it may not be sufficient for accepting the alternate hypothesis. The logarithmic Bayesian factor is also out of line

Mortality rate Bayesian efficacy thresholds				
Study	95%	log Bayes = 2 thresholds		
	threshold	$p_2 = 2\%$	$p_2 = 5\%$	$p_2 = 10\%$
DSZ study	3.8%	N/A	N/A	3.9%
Zelenko April 2020	1.8%	N/A	1.8%	1.5%
Zelenko June 2020	1.0%	1.0%	0.8%	0.6%
Procter I	1.7%	N/A	1.9%	1.3%
Procter II	0.84%	0.87%	0.7%	0.6%
Raoult	0.79%	0.82%	< 0.7%	< 0.7%
Hospitalization rate Bayesian efficacy thresholds				
Study	95%	log Bayes = 2 thresholds		
	threshold	$p_2 = 10\%$	$p_2 = 15\%$	$p_2 = 20\%$
DSZ study	7.0%	9.5%	7.8%	7.2%
Zelenko April 2020	3.2%	3.2%	3.0%	2.9%
Zelenko June 2020	2.7%	2.6%	2.5%	2.4%
Procter I	4.1%	4.3%	4.0%	3.7%
Procter II	3.6%	3.7%	3.5%	3.4%

FIG. 14: Comparison of the 95% confidence efficacy threshold (Sterne interval [68]) for mortality and hospitalization rate reduction with the Bayes factor efficacy thresholds at log Bayes = 2, using maximum untreated mortality rate p_2 for high risk patients at $p_2 \in \{0.02, 0.05, 0.10\}$ and maximum untreated hospitalization rate p_2 for high risk patients at $p_2 \in \{0.10, 0.15, 0.20\}$, for the DSZ study treatment group [2], Zelenko's complete April 2020 data set [13], Zelenko's complete June 2020 data set [14], Procter's observational studies [15, 16], and Raoult's high risk (older than 60) treatment group [17].

Bayes factors at the mortality rate efficacy thresholds				
Study	99% threshold	$p_2 = 0.02$	$p_2 = 0.05$	$p_2 = 0.1$
DSZ study	5.3%	N/A	N/A	2.70
Zelenko April	2.4%	N/A	2.81	3.27
Zelenko June	1.2%	2.53	3.21	3.57
Procter I	2.3%	N/A	2.72	3.17
Procter II	1.08%	2.55	3.17	3.53
Raoult	0.96%	2.57	3.16	3.51
Bayes factors at the hospitalization rate efficacy thresholds				
Study	99% threshold	$p_2 = 0.02$	$p_2 = 0.05$	$p_2 = 0.1$
DSZ study	8.8%	1.83	2.42	2.67
Zelenko April	3.9%	2.75	3.00	3.17
Zelenko June	3.0%	2.77	3.00	3.16
Procter I	4.9%	2.55	2.85	3.02
Procter II	4.0%	2.63	2.89	3.05

FIG. 15: Bayes factor (decimal logarithm) corresponding to the 99% efficacy threshold (Sterne interval [68]) for mortality and hospitalization rate reduction, using maximum untreated mortality rate p_2 for high risk patients at $p_2 \in \{0.02, 0.05, 0.10\}$ and maximum untreated hospitalization rate p_2 for high risk patients at $p_2 \in \{0.10, 0.15, 0.20\}$, for the DSZ study treatment group [2], Zelenko's complete April 2020 data set [13], Zelenko's complete June 2020 data set [14], Procter's observational studies [15, 16], and Raoult's high risk (older than 60) treatment group [17].

for the Procter I case series [15] for hospitalization rate reduction with $p_2 = 10\%$, however this is not a serious concern since that particular choice of p_2 is unrealistically small.

In Fig. 14, we compare the efficacy thresholds for rejecting the null hypothesis with the corresponding 95% confidence Bayesian thresholds, obtained by the inequality $b(x_0, p_2) \geq 2$ for accepting the alternate hypothesis. For the DSZ study [2], we see that the corresponding Bayesian thresholds for hospitalization rate reduction range from 7.2% to 9.5%, which lie above the 7.0% threshold obtained via the p-value. So, the most cautious course of action is to opt for the 9.5% threshold, which is still below most of our estimates for hospitalization probability of untreated patients. For the DSZ study [2], for both $p_2 = 2\%$ and $p_2 = 5\%$, the logarithmic Bayesian factor for mortality rate reduction does not go above the decisive threshold for any value of x with $a/N \leq x \leq p_2$, consequently the corresponding Bayesian thresholds are undefined, and for $p_2 = 10\%$ we find a Bayesian mortality rate reduction threshold of 3.9% which is slightly larger than the p-value threshold of 3.8%. For the Procter I case series [15], there is a weak indication that the 4.1% efficacy threshold for hospitalization rate reduction might have to be increased to 4.3%, and the mortality rate reduction threshold increased from 1.7% to 1.9%. Likewise for the Procter II case series [16], an increase of the hospitalization rate reduction efficacy threshold from 3.6% to 3.7% is weakly indicated. Both adjustments are negligible and inconsequential. For the Zelenko April 2020 [13] and Zelenko June 2020 [14] case series, where the sample sizes are much larger, we see that the overall trend is for the Bayesian thresholds to be far more lenient than the ones obtained via the p-value. This is possibly attributed to a very strong signal of efficacy in the data.

It is interesting to repeat the Bayesian analysis on the efficacy thresholds for mortality rate reduction and hospitalization rate reduction for 99% confidence and 99.9% confidence. We have seen that the Bayesian adjustments to the 95% confidence efficacy thresholds, when they are needed, are very small, so the relevant question is whether this pattern continues when the demanded confidence increases to 99% or 99.9%. Fig. 15 and Fig. 16 show the values of the logarithmic Bayesian factor $b_2(x_0, p_2)$ at the mortality and hospitalization efficacy thresholds for 99% and 99.9% confidence, as determined solely from the p-value, and for various values of p_2, as previously discussed. Note that for Fig. 15 the decisive Bayesian factor threshold corresponding to 99% confidence is $b_2(x_0, p_2) \geq 2.7$. Likewise, in Fig. 16, the decisive Bayesian factor threshold corresponding to 99.9% confidence is $b_2(x_0, p_2) \geq 3.7$. We see that for the most part the logarithmic Bayesian factors are either above or near their respective thresholds.

Likewise, in Fig. 17 and Fig. 18 we are comparing the mortality and hospitalization rate reduction efficacy thresholds determined via the p-value, against the corresponding efficacy thresholds determined using the logarithmic Bayesian factor $b_2(x_0, p_2)$, for 99% and 99.9% confidence correspondingly.

Bayes factors at the mortality rate efficacy thresholds				
Study	99.9% threshold	$p_2 = 0.02$	$p_2 = 0.05$	$p_2 = 0.1$
DSZ study	7.0%	N/A	N/A	3.51
Zelenko April 2020	2.9%	N/A	3.47	4.00
Zelenko June 2020	1.6%	3.43	4.34	4.73
Procter I	3.1%	N/A	3.59	4.16
Procter II	1.4%	3.38	4.15	4.53
Raoult	1.18%	3.49	4.16	4.52
Bayes factors at the hospitalization rate efficacy thresholds				
Study	99.9% threshold	$p_2 = 0.02$	$p_2 = 0.05$	$p_2 = 0.1$
DSZ study	10.6%	N/A	3.17	3.49
Zelenko April 2020	4.7%	3.68	3.97	4.15
Zelenko June 2020	3.5%	3.75	4.00	4.16
Procter I	5.9%	3.45	3.80	3.99
Procter II	4.5%	3.54	3.82	3.99

FIG. 16: Bayes factor (decimal logarithm) corresponding to the 99.9% efficacy threshold (Sterne interval [68]) for mortality and hospitalization rate reduction, using maximum untreated mortality rate p_2 for high risk patients at $p_2 \in \{0.02, 0.05, 0.10\}$ and maximum untreated hospitalization rate p_2 for high risk patients at $p_2 \in \{0.10, 0.15, 0.20\}$, for the DSZ study treatment group [2], Zelenko's complete April 2020 data set [13], Zelenko's complete June 2020 data set [14], Procter's observational studies [15, 16], and Raoult's high risk (older than 60) treatment group [17].

Mortality rate Bayesian efficacy thresholds				
Study	99%	log Bayes = 2.7 thresholds		
	threshold	$p_2 = 2\%$	$p_2 = 5\%$	$p_2 = 10\%$
DSZ study	5.3%	N/A	N/A	5.3%
Zelenko April 2020	2.4%	N/A	2.4%	2.0%
Zelenko June 2020	1.2%	1.3%	1.1%	0.9%
Procter I	2.3%	N/A	2.3%	1.9%
Procter II	1.08%	1.14%	0.92%	0.80%
Raoult	0.96%	1.0%	0.86%	0.77%
Hospitalization rate Bayesian efficacy thresholds				
Study	99%	log Bayes = 2.7 thresholds		
	threshold	$p_2 = 10\%$	$p_2 = 15\%$	$p_2 = 20\%$
DSZ study	8.8%	N/A	9.5%	8.9%
Zelenko April 2020	3.9%	N/A	3.7%	3.5%
Zelenko June 2020	3.0%	3.0%	2.9%	2.8%
Procter I	4.9%	5.1%	4.8%	4.6%
Procter II	4.0%	4.1%	3.9%	3.8%

FIG. 17: Comparison of the 99% confidence efficacy threshold (Sterne interval [68]) for mortality and hospitalization rate reduction with the Bayes factor efficacy thresholds at log Bayes = 2.7, using maximum untreated mortality rate p_2 for high risk patients at $p_2 \in \{0.02, 0.05, 0.10\}$ and maximum untreated hospitalization rate p_2 for high risk patients at $p_2 \in \{0.10, 0.15, 0.20\}$, for the DSZ study treatment group [2], Zelenko's complete April 2020 data set [13], Zelenko's complete June 2020 data set [14], Procter's observational studies [15, 16], and Raoult's high risk (older than 60) treatment group [17].

We see that the Bayesian perturbations to the efficacy thresholds are for the most part negligible for both 99% and 99.9% confidence, continuing the similar pattern that we have observed for the 95% confidence efficacy thresholds.

As a practical matter, in all cases, we want to see both the frequentist and Bayesian thresholds exceeded before claiming a statistically significant result. For the case series under consideration, we see that the Bayesian adjustments to the efficacy thresholds for mortality and hospitalization rate reduction are negligible and therefore do not impact the analysis of the preceding sections.

5. DISCUSSION AND CONCLUSIONS

Our findings fully support risk stratification in the management of acute COVID-19, with the intent of reducing the intensity and duration of symptoms and by that mechanism, lower the risk of hospitalization and death. Although COVID-19 is generally known as a respiratory disease, there is an accumulation of evidence [37, 89, 90] that it is also, if not primarily, a vascular disease, with endothelial injury having a major role in sustained permanent injuries, hospitalizations, and death. The most impactful countermeasure that can prevent these adverse outcomes is early outpatient treatment, using multiple drugs in combination, that stops viral replication at the first phase of the illness, and mitigates the injuries caused by the hyper inflammatory COVID-19 pneumonia phase and the subsequent thromboembolic phase.

Looking back, after 2 years of going through the COVID-19 pandemic, one of the lessons learned is that some of the key discoveries for the successful treatment of a novel disease emerge from the experience of the frontline doctors that are directly confronted with the need to find a way to help their patients. Although the orthodox approach is to consider possible treatments as unproven until they are validated with an RCT, in real life, it is possible to be confronted with a situation where the observational data is sufficiently strong to justify the immediate adoption of a treatment protocol, and to raise the ethical concern of whether it is appropriate to even conduct the RCT, and deny treatment to a very large cohort of patients, in order to form a control group. Consequently, there is a need to be able to analyze observational data in a statistically rigorous way.

We have provided a hybrid statistical framework for assessing observational evidence that combines both frequentist and Bayesian methods; the frequentist methods aim to control the p-value for rejecting the null hypothesis, whereas the Bayesian methods aim to control the false positive rate. The two methods are complementary and not mutually exclusive. The core mathematical ideas are very simple, however there are some subtleties concerning the relationship of the frequentist methods with the exact Fisher test and the binomial proportion confidence interval problem. We have also discussed some counterintuitive misbehavior of the p-value, that we no-

Mortality rate Bayesian efficacy thresholds				
Study	99.9%	log Bayes = 3.7 thresholds		
	threshold	$p_2 = 2\%$	$p_2 = 5\%$	$p_2 = 10\%$
DSZ study	7.0%	N/A	N/A	7.4%
Zelenko April	2.9%	N/A	3.1%	2.7%
Zelenko June	1.6%	1.8%	1.4%	1.3%
Procter I	3.1%	N/A	3.2%	2.8%
Procter II	1.4%	1.53%	1.26%	1.14%
Raoult	1.18%	1.23%	1.08%	1.01%
Hospitalization rate Bayesian efficacy thresholds				
Study	99.9%	log Bayes = 3.7 thresholds		
	threshold	$p_2 = 10\%$	$p_2 = 15\%$	$p_2 = 20\%$
DSZ study	10.6%	N/A	11.9%	11.1%
Zelenko April 2020	4.7%	4.8%	4.5%	4.4%
Zelenko June 2020	3.5%	3.5%	3.4%	3.3%
Procter I	5.9%	6.2%	5.8%	5.7%
Procter II	4.5%	4.6%	4.5%	4.4%

FIG. 18: Comparison of the 99.9% confidence efficacy threshold (Sterne interval [68]) for mortality and hospitalization rate reduction with the Bayes factor efficacy thresholds at log Bayes = 3.7, using maximum untreated mortality rate p_2 for high risk patients at $p_2 \in \{0.02, 0.05, 0.10\}$ and maximum untreated hospitalization rate p_2 for high risk patients at $p_2 \in \{0.10, 0.15, 0.20\}$, for the DSZ study treatment group [2], Zelenko's complete April 2020 data set [13], Zelenko's complete June 2020 data set [14], Procter's observational studies [15, 16], and Raoult's high risk (older than 60) treatment group [17].

ticed during the course of this investigation, that is inherent to the definition of the p-value itself. Specifically we've noticed that the p-value does not have a precisely monotonic relationship with our intuitive understanding of statistical confidence in rejecting the null hypothesis. Empirically, we have noticed that this misbehavior intensifies with small sample sizes, and is diminished with increasing sample sizes [22].

The main weakness of the proposed statistical methodology is that it has to be limited only to the assessment of treatments that are based on repurposed medications [60] with a known excellent safety record. It would be highly inappropriate to use this approach on new medications, or other countermeasures, where the balance of risks and benefits is yet to be determined. Furthermore, the analysis of the treatment group case series needs to be compared with a model that can at minimum lower-bound the probability of adverse outcomes without treatment, based on our prior knowledge. On the other hand, the development of this model can be done independently from the analysis of the treatment group case series.

One way in which our approach deviates from the usual way of doing things, is that we are using the proposed statistical methodology to assess the efficacy of the entire treatment algorithm against supportive care. Both of the original Zelenko protocol [2] and the more enhanced McCullough protocol [18–20] are examples of sequenced multi-drug treatment protocols. Furthermore, both protocols are algorithmic, in the sense that treatment is customized to the individual patient, based on the patient's medical history and the response to treatment. For the case of the Zelenko protocol [2] this is done via the risk stratification of patients to low-risk and high-risk patients. For the case of the McCulloch protocol [18–20], this is done both by risk stratification and also by accounting for the progression of the illness through the three distinct stages and response to treatment. Consequently, the immediate goal is not to establish that any particular drug is effective. The goal is to establish that the treatment algorithm itself is effective, so that it can be deployed rapidly on an emergency basis and be subsequently improved over time with further research.

A possible theoretical criticism is that the particular case series that we have considered may have selection bias, i.e. they were published and came to our attention as a result of their positive outcomes, and not because they are representative of the typical outcomes of most treatment centers that employ early outpatient treatment protocols against COVID-19. This is mitigated to some extent by the fact that we have reported case series from three different treatment centers, two in the United States and one in France, with consistent mortality rates, therefore this consistency is compelling statistical evidence against selection bias. More importantly, for both of the Zelenko [2, 13, 14] and Procter [15, 16] case series, we have two consecutive reports over two consecutive time intervals replicating the hospitalization and mortality rate reduction outcomes, and these replications are additional statistical evidence against a theoretical selection bias.

The case series that we have analyzed in this paper add up to a total of 3164 high-risk patients. It is currently estimated that the total number of high-risk patients that have been treated with early outpatient treatment protocols throughout the United States may exceed this number by one or two orders of magnitude [48]. Unfortunately, no resources have been allocated to study this data by our public health agencies, but we can make some suggestions about how such an analysis could be carried out. One idea for quickly analyzing a very large dataset is to extract the age > 50 and/or age > 65 part of the database, calculate the corresponding efficacy thresholds for hospitalization rate reduction and mortality rate reduction, and compare them with the CDC estimates [56–58] for number of hospitalizations and deaths for these age groups over the total number of cases with symptomatic illness. Given a large enough data set, it would also be interesting to risk-stratify the age > 50 and/or age > 65 cohorts further with respect to number of days between initial symptoms and initiation of treatment and calculate the efficacy thresholds as a function of the delay in initiating treatment. Furthermore, it would be useful to breakdown the case series data in sequential time intervals corresponding to different waves and different variants of the SARS-CoV-2 virus. This analysis would inadvertently not include younger patients that are high risk due to comorbidities or shortness of breath presentation, however it has the advantage that it can be carried out quickly with limited resources. Analyzing the data from several more treatment centers, that have adopted early outpatient treatment protocols for high-risk patients would further mitigate the potential for selection bias.

With substantial resources, a more detailed analysis is possible that can consider the entire dataset and actually estimate the treatment efficacy. Given a case series of N patients, one can input the medical history of each patient to the Cleveland Clinic calculator [87] and use their mathematical model to predict the probability of hospitalization and death for each patient individually. Knowing the corresponding sequence of probabilities $\mathbf{q} = (p_1, p_2, \ldots, p_N)$ for an adverse outcome (hospitalization or death) for all patients, the probability $\mathrm{pr}(N, a|\mathbf{q})$ of seeing a adverse outcomes follows a Poison binomial distribution [91], and it can be substituted to Eq. (2) in order to calculate the p-value for rejecting the null hypothesis of no treatment efficacy. Because the probability of an adverse outcome is known for each patient, note that there is no need to worry about calculating any efficacy thresholds, and it is possible instead to directly calculate the p-value for rejecting the null hypothesis.. Furthermore, since the mean of the Poison binomial distribution is the average $q = (1/N)(p_1 + p_2 + \ldots + p_n)$ of the individual probabilities, one can calculate the risk ratio via the equation $\mathrm{RR} = a/(qN)$. To conduct the corresponding Bayesian analysis, we can assume that the effect of the early outpatient treatment is to reduce the probabilities of adverse outcome by a numerical factor x to $x\mathbf{q} = (xp_1, xp_2, \ldots, xp_N)$ with $0 \leq x \leq 1$ and use the Poison binomial distribution $\mathrm{pr}(N, a|x\mathbf{q})$ in Eq. (35) and Eq. (38) to calculate the corresponding integrals needed for the Bayesian factor. All other aspects of the Bayesian analysis would remain the same, except that the hypothesis being validated would not concern any efficacy thresholds but it would instead concern hypotheses about the actual efficacy x of the early outpatient treatment protocol.

That said, we do not mean to imply that such a detailed analysis is necessary in order to greenlight the use of the investigated early outpatient treatment protocols for COVID-19. However, the fact that such a detailed analysis is possible to carry out, using existing data and prior mathematical modeling, is highly relevant with respect to assessing the ethics of validating the McCullough protocol at this time, using an RCT. A limitation of the Cleveland Clinic calculator is that it should ideally be used in conjunction with case series over time intervals that are aligned with the data set used to train the calculator's mathematical predictive model. Because the Cleveland Clinic calculator used data collected between March 4th 2020 and July 14th 2020 it can certainly be applied to case series up until July 2020. However we believe that it can also be extended up until and including the Delta variant, that became dominant towards the end of 2021, since all of these subsequent variants were just as hard to treat or harder than the initial waves in 2020.

Notwithstanding economically-motivated obstacles [86, 92] that have been placed against the adoption of early treatment protocols for COVID-19, everything that we have been through during the last two years vindicates the position of Frieden [50] that there is an urgent need to leverage and overcome the limitations of real-world evidence data, in order to deploy a timely life-saving response to urgent health issues. There is still an opportunity to learn much by analyzing data from various treatment centers here in the United States that treated COVID-19 with early outpatient treatment protocols, as well as treatment centers from all around the world. More importantly, it is necessary to reflect on and develop policies and procedures for leveraging the direct experience of frontline doctors treating patients, towards an agile and effective response to future epidemics and pandemics.

Author contributions

EG conceptualized the mathematical framework, analyzed the case series data, and wrote the first draft of the paper. PAC and VZ contributed to the literature review, as well as writing and editorial changes to the manuscript. VZ contributed copies of his three public letters [13, 14, 21], attached to our supplementary material document [22], containing some of the consecutive case series data analyzed in the manuscript. VZ also contributed data on how his treatment protocol evolved over time during 2020. All authors read and approved the final version of the manuscript.

Ethics approval and consent to participate

Not applicable.

Acknowledgements

It is a pleasure to thank Roland Derwand and Harvey Risch for correspondence and encouragement. In particular, we wish to thank Roland Derwand for bringing the Israeli study [55] to our attention, and Harvey Risch for highlighting the importance of the paper by Frieden [50] and the paper by Deaton and Cartwright [51]. We also wish to acknowledge Lawrence Huntoon for encouraging us to look into a Bayesian approach, at a very early stage of this research project.

Funding

The research has received no external funding.

Conflict of interest

The authors declare no conflicts of interest. Peter A. McCullough is serving as the editor-in-chief of this journal.

Appendix A: Exact Fisher test in the limit of an infinite control group

Let N be the total number of patients in the treatment group, let a be the number of patients with an adverse outcome (hospitalization or death) in the treatment group, let M be the total number of patients in the control group, and let b be the number of patients in the control group with an adverse

outcome. In this appendix we will show that in the limit of an infinite control group (M, b) with $x = b/M$, the p-value $p(N, a, M, b)$ obtained from the two-tail exact Fisher test converges to $p(N, a, x)$.

In the exact Fisher test, we assume that N, M, and $a+b$, are fixed numbers, and under the null hypothesis, we also assume that the distribution of the total $a + b$ patients with an adverse outcome between the treatment group and control group is random, with equal probability for every possible combination. It follows that under the null hypothesis, the probability of seeing a particular event (N, a, M, b) is given by

$$\mathrm{pr}(N, a, M, b) = \frac{\binom{a+b}{b}\binom{N+M-a-b}{N-a}}{\binom{N+M}{N}}. \tag{A1}$$

The corresponding p-value is the probability of observing the event (N, a, M, b) or any other less probable event, and it is given by

$$\mathrm{pr}(N, a, M, b) = \sum_{n=0}^{\min\{N, a+b\}} \mathrm{pr}(N, n, M, a+b-n) H(\mathrm{pr}(N, a, M, b) - \mathrm{pr}(N, n, M, a+b-n)), \tag{A2}$$

We note that the summation variable n is restricted by both the total size N of the treatment group and the total number $a + b$ of the patients with an adverse outcome, so the permissible range for all possible events is $0 \le n \le \min\{N, a + b\}$.

A key insight is that in the definition of $\mathrm{pr}(N, a, M, b)$, the variable M can be replaced with a continuous real number, because it appears only in the top argument of the corresponding binomial coefficients. Recall that for all $a \in \mathbb{R}$ and $n \in \mathbb{N}$ the extended definition of the binomial coefficient is given by

$$\binom{a}{n} = \frac{1}{n!}\prod_{\lambda=1}^{n}(a+1-\lambda) = \frac{1}{n!}\prod_{\lambda=1}^{n}(a+1-(n-\lambda+1)) = \frac{1}{n!}\prod_{\lambda=1}^{n}(a-n+\lambda). \tag{A3}$$

On the second step we have used the transformation $\lambda \mapsto n - \lambda + 1$ which effectively reverses the order of factors in the product. It follows that for all $M \in \mathbb{R}$ the corresponding M-dependent binomial coefficients are given by

$$\binom{N+M-a-b}{N-a} = \frac{1}{(N-a)!}\prod_{\lambda=1}^{N-a}((N+M-a-b)-(N-a)+\lambda) = \frac{1}{(N-a)!}\prod_{\lambda=1}^{N-a}(M-b+\lambda), \tag{A4}$$

and

$$\binom{N+M}{N} = \frac{1}{N!}\prod_{\lambda=1}^{N}((N+M)-N+\lambda) = \frac{1}{N!}\prod_{\lambda=1}^{N}(M+\lambda), \tag{A5}$$

and thus, the hypergeometric probability distribution $\mathrm{pr}(N, a, M, b)$ can be rewritten as

$$\mathrm{pr}(N, a, M, b) = \frac{\binom{a+b}{b}\binom{N+M-a-b}{N-a}}{\binom{N+M}{N}} \tag{A6}$$

$$= \frac{(a+b)!}{a!b!}\left[\frac{1}{(N-a)!}\prod_{\lambda=1}^{N-a}(M-b+\lambda)\right]\left[N!\prod_{\lambda=1}^{N}\left(\frac{1}{M+\lambda}\right)\right] \tag{A7}$$

$$= \frac{N!}{a!(N-a)!}\frac{(a+b)!}{b!}\prod_{\lambda=1}^{N-a}(M-b+\lambda)\prod_{\lambda=1}^{N}\left(\frac{1}{M+\lambda}\right) \tag{A8}$$

$$= \binom{N}{a}\prod_{\lambda=1}^{a}(b+\lambda)\prod_{\lambda=1}^{N-a}(M-b+\lambda)\prod_{\lambda=1}^{a}\left(\frac{1}{M+\lambda}\right)\prod_{\lambda=1}^{N-a}\left(\frac{1}{M+a+gl}\right) \tag{A9}$$

$$= \binom{N}{a} \prod_{\lambda=1}^{a} \left(\frac{b+\lambda}{M+\lambda} \right) \prod_{\lambda=1}^{N-a} \left(\frac{M-b+\lambda}{M+a+\lambda} \right). \tag{A10}$$

To take the limit of an infinite control group with probability x of an adverse outcome, we set $b = xM$, or equivalently $M = (1/x)b$, and take a sequence limit $b \in \mathbb{N}$ to infinity. We conclude that

$$\lim_{b\in\mathbb{N}} \mathrm{pr}(N, a, (1/x)b, b) = \binom{N}{a} \left[\prod_{\lambda=1}^{a} \lim_{b\in\mathbb{N}} \left(\frac{b+\lambda}{(1/x)b+\lambda} \right) \right] \left[\prod_{\lambda=1}^{N-a} \lim_{b\in\mathbb{N}} \left(\frac{(1/x)b-b+\lambda}{(1/x)b+a+\lambda} \right) \right] \tag{A11}$$

$$= \binom{N}{a} \left(\frac{1}{1/x} \right)^{a} \left(\frac{1/x-1}{1/x} \right)^{N-a} \tag{A12}$$

$$= \binom{N}{a} x^{a}(1-x)^{N-a} = \mathrm{pr}(N, a|x). \tag{A13}$$

An immediate consequence is that the corresponding p-values satisfy a similar relationship that reads

$$\lim_{b\in\mathbb{N}^*} p(N, a, (1/x)b, b) = p(N, a|x). \tag{A14}$$

The probability sums on both sides of Eq. (A14) involve a variable n that goes from 0 to N, making the number of terms on the left-hand-side probability sum independent of the size of the control group, as soon as b is large enough. This makes it possible to derive Eq. (A14) as an immediate consequence of Eq. (A13).

Appendix B: Monotonicity of the Bayesian factor

We prove that the function $b_0(x_0, p_2, t)$ is initially increasing and then decreasing with respect to t with a maximum in the interval $[a/N, 1]$. We recall that

$$b_0(x_0, p_2, t) = \log \left[\frac{p_2 - x_0}{t} \frac{\int_0^t x^a(1-x)^{N-a} \mathrm{d}x}{\int_{x_0}^{p_2} x^a(1-x)^{N-a} \mathrm{d}x} \right], \tag{B1}$$

consequently maximizing the function $b_0(x_0, p_2, t)$ is equivalent to maximizing

$$g(t) = \frac{1}{t} \int_0^t x^a(1-x)^{N-a} \, \mathrm{d}x, \tag{B2}$$

since all other factors are independent of t. For our argument, it is simpler to work with the more abstract definition

$$g(t) = \frac{1}{t} \int_0^t f(x) \, \mathrm{d}x, \tag{B3}$$

and assume that the function $f(x)$ is increasing in the interval $[0, a/N]$, decreasing in the interval $[a/N, 1]$, and also satisfies $f(1) = 0$ and $f(x) > 0$ for all $x \in (0, 1)$. These are all general assumptions that are indeed satisfied by the binomial distribution $f(x) = x^a(1-x)^{N-a}$. Differentiating with respect to t gives

$$g'(t) = \frac{-1}{t^2} \int_0^t f(x) \, \mathrm{d}x + \frac{f(t)}{t}. \tag{B4}$$

From the assumptions $f(1) = 0$ and $f(x) > 0$ for all $x \in (0, 1)$, it immediately follows that

$$g'(1) = -\int_0^1 f(x) \, \mathrm{d}x < 0. \tag{B5}$$

Next, we apply the integral mean-value theorem on the interval $[0, a/N]$ which requires the assumption that $f(x) > 0$ for all $x \in (0, a/N]$ and it follows that there exists $\xi \in [0, a/N]$ such that

$$f(\xi) = \frac{1}{a/N} \int_0^{a/N} f(x) \, \mathrm{d}x. \tag{B6}$$

We use this equation to show that

$$g'(a/N) = \frac{-1}{(a/N)^2} \int_0^{a/N} f(x) \, \mathrm{d}x + \frac{f(a/N)}{a/N} \tag{B7}$$

$$= \frac{-f(\xi)}{a/N} + \frac{f(a/N)}{a/N} \tag{B8}$$

$$= \frac{(f(a/N) - f(\xi))N}{a} > 0. \tag{B9}$$

Here, the inequality step is justified by the assumption that the function $f(x)$ is increasing in the interval $[0, a/N]$. It follows via the Bolzano theorem that there is at least one $t_0 \in [a/N, 1]$ such that $g'(t_0) = 0$, making all such t_0 critical points that are the possible local minimum or maximum points of $g(t)$. From Eq. (B4), it follows that all such critical points t_0 also satisfy the equation

$$f(t_0) = \frac{1}{t_0} \int_0^{t_0} f(x) \, \mathrm{d}x. \tag{B10}$$

We shall now use the second derivative test to show that any such critical points have to be local maxima, which in turn implies the uniqueness of only one such local maximum point in the interval $[a/N, 1]$. The second derivative of the function $g(t)$ is given by

$$g''(t) = \frac{\mathrm{d}}{\mathrm{d}t} \left[\frac{-1}{t^2} \int_0^t f(x) \, \mathrm{d}x + \frac{f(t)}{t} \right] \tag{B11}$$

$$= \frac{2}{t^3}\int_0^t f(x)\,\mathrm{d}x - \frac{f(t)}{t^2} - \frac{f(t)}{t^2} + \frac{f'(t)}{t} \quad \text{(B12)}$$

$$= \frac{2}{t^3}\int_0^t f(x)\,\mathrm{d}x - \frac{2f(t)}{t^2} + \frac{f'(t)}{t}, \quad \text{(B13)}$$

and for $t = t_0$, it follows that

$$g''(t_0) = \frac{2}{t_0^3} t_0 f(t_0) - \frac{2f(t_0)}{t_0^2} + \frac{f'(t_0)}{t_0} = \frac{f'(t_0)}{t_0} < 0. \quad \text{(B14)}$$

Here, the last inequality step is justified by the assumption that the function $f(x)$ is decreasing over the interval $[a/N, 1]$ and furthermore that $t_0 \in [a/N, 1]$. We conclude that all critical points in the interval $[a/N, 1]$ have to be local maxima, and by necessity this means that only one such local maximum actually exists in the interval $[a/N, 1]$. This concludes the proof of our claim.

[1] Gautret P., Lagier J.C., Parola P., et al. Hydroxychloroquine and azithromycin as a treatment of COVID-19: Results of an open-label non-randomized clinical trial *International Journal of Antimicrobial Agents.* 2020;56:105949.

[2] Scholz M., Derwand R., Zelenko V.. COVID-19 outpatients - Early risk-stratified treatment with zinc plus low dose hydroxychloroquine and azithromycin: A retrospective case series study. *International Journal of Antimicrobial Agents.* 2020;56:106214. https://doi.org/10.1016/j.ijantimicag.2020.106214.

[3] Chetty S.. Elucidating the pathogenesis and Rx of COVID reveals a missing element *Modern Medicine.* 2020;45 (5):28-31.

[4] Marik P.E., Kory P., Varon J., Iglesias J., Meduri G.U.. MATH+ protocol for the treatment of SARS-CoV-2 infection: the scientific rationale *Expert Review of Anti-infective Therapy.* 2021;19(2):129-135.

[5] Sherman R.E., Anderson S.A., Pan G.J.D., et al. Real-World Evidence - What Is It and What Can It Tell Us? *New England Journal of Medicine.* 2016;375:2293-2297.

[6] Wasserstein A.L., Lazar N.A.. The ASA's statement on p-values: context, process, and purpose *The American Statistician.* 2016;70 (2):129-133.

[7] Berger J.I.. Could Fisher, Jeffreys, and Neyman have agreed on testing? *Statistical Science.* 2003;18:1-32.

[8] Goodman S.. Toward evidence-based medical statistics. 1: The P value fallacy *Annals of Internal Medicine.* 1999.;130 (12):995-1004.

[9] Goodman S.. Toward evidence-based medical statistics. 2: The Bayes factor *Annals of Internal Medicine.* 1999.;130 (12):1005-1013.

[10] Morey R.D., Romeijn J.W, Rouder J.N.. The philosophy of Bayes factors and the quantification of statistical evidence *Journal of Mathematical Psychology.* 2016.;72:6-18.

[11] Kass R.E., Raftery A.E.. Bayes Factors *Journal of the American Statistical Association.* 1995.;90 (430):791.

[12] Colquhoun D.. The False Positive Risk: A Proposal Concerning What to Do About p-Values *The American Statistician.* 2019;73:sup1:192-201.

[13] Zelenko V.. To all medical professionals around the world [letter]. April 28, 2020. Available at: https://tinyurl.com/yxk8ssco. Accessed November 26, 2020, now censored by Google. Attached to supplementary material.

[14] Zelenko V.. To Dr. Moshe Bar Siman Tov [letter]. June 14, 2020. Available at: https://tinyurl.com/y4hw7dye. Accessed November 26, 2020, now censored by Google. Attached to supplementary material.

[15] Procter B.C., Ross C., Pickard V., Smith E., Hanson C., McCullough P.A.. Clinical outcomes after early ambulatory multidrug therapy for high-risk SARS-CoV-2 (COVID-19) infection *Reviews in Cardiovascular Medicine.* 2021;21:611-614.

[16] Procter B.C., Ross C., Pickard V., Smith E., Hanson C., McCullough P.A.. Early Ambulatory Multidrug Therapy Reduces Hospitalization and Death in High-Risk Patients with SARS-CoV-2 (COVID-19) *International Journal of Innovative Research in Medical Science.* 2021;6:219-221.

[17] Million M., Lagier J.C., Tissot-DuPont H., et al. Early Treatment with Hydroxychloroquine and Azithromycin in 10,429 COVID-19 Outpatients: A Monocentric Retrospective Cohort Study *Reviews in Cardiovascular Medicine.* 2021;22:1063-1072.

[18] McCullough P.A., Kelly R.J., Ruocco G., et al. Pathophysiological Basis and Rationale for Early Outpatient Treatment of SARS-CoV-2 (COVID-19) Infection. *The American Journal of Medicine.* 2020;134:16-22.

[19] McCullough P.A.. Innovative Early Sequenced Multidrug Therapy for SARS-CoV-2 (COVID-19) Infection to Reduce Hospitalization and Death *International Journal of Medical Science and Clinical invention.* 2020;7:5139-5150.

[20] McCullough P.A., Alexander P.E., Armstrong R., et al. Multifaceted highly targeted sequential multidrug treatment of early ambulatory high-risk SARS-CoV-2 infection (COVID-19) *Reviews in Cardiovascular Medicine.* 2020;21 (4):517-530.

[21] Zelenko V.. Correspondence from Dr Vladimir Zelenko on Treatment of COVID-19 in New York. March 23, 2020. Available at: https://tinyurl.com/yx2cxf6q. Accessed November 26, 2020. Attached to supplementary material.

[22] Gkioulekas E.. Supplementary material: Frequentist and Bayesian analysis methods for case series data and application to early outpatient COVID-19 treatment case series 2022.

[23] Derwand R., Scholz M.. Does zinc supplementation enhance the clinical efficacy of chloroquine/hydroxychloroquine to win today's battle against COVID-19? *Medical Hypotheses.* 2020;142:109815.

[24] Wessels I., Rolles B., Rink L.. The Potential Impact of Zinc Supplementation on COVID-19 Pathogenesis *Frontiers in Immunology.* 2020.;11:1712.

[25] Stricker R.B., Fesler M.C.. A novel plan to deal with SARS-CoV-2 and COVID-19 disease *Journal of Medical Virology.* 2020;92:1394-1395.

[26] Merritt L.D.. The Treatment of Viral Diseases: Has the Truth Been Suppressed For Decades? *Journal of the American Physicians and Surgeons.* 2020;25 (3):79-82.

[27] Galvez J., Zanni R., Galvez-Llompart M., Benlloch J.M.. Macrolides May Prevent Severe Acute Respiratory Syndrome Coronavirus 2 Entry into Cells: A Quantitative Structure Activity Relationship Study and Experimental Validation *Journal of Chemical Information and Modeling.* 2021;61:2016-2025.

[28] Heras E., Garibaldi P., Boix M., et al. COVID-19 mortality risk factors in older people in a long-term care center *European Geriatric Medicine.* 2021;12:601-607.

[29] Vincent M.J., Bergeron E., Benjannet S., et al. Chloroquine is a potent inhibitor of SARS coronavirus infection and spread *Virology Journal.* 2005;2:69.

[30] Burrows W.F.. The abortive treatment of influenza with quinine dihydrochloride *Medical Record.* 1918;94:1081-1082.

[31] Velthuis A.J. Te, Fodor E.. Influenza virus RNA polymerase: insights into the mechanisms of viral RNA synthesis *Nature Reviews Microbiology.* 2016;14(8):479-493.

[32] Biancatelli1 R.M.L.C., Berrill M., Catravas J.D., Marik P.E.. Quercetin and Vitamin C: An Experimental, Synergistic Therapy for the Prevention and Treatment of SARS-CoV-2 Related Disease (COVID-19). *Frontiers in Immunology.* 2020;11:1451.

[33] Dabbagh-Bazarbachi H., Clergeaud G., Quesada I.M., Ortiz M., O'Sullivan C.K., Fernandez-Larrea J.B.. Zinc Ionophore Activity of Quercetin and Epigallocatechin-gallate: From Hepa 1-6 Cells to a Liposome Model *Journal of Agricultural and Food Chemistry.* 2014;62 (32):8085-8093.

[34] Balakrishnan A., Price E., Luu C., et al. Biochemical Characterization of Respiratory Syncytial Virus RNA-Dependent RNA Polymerase Complex *ACS Infectious Diseases.* 2020;6:2800-2811.

[35] Yu D.S., Weng T.H., Wu X.X., et al. The lifecycle of the Ebola virus in host cells *Oncotarget.* 2017;8(33):55750-55759.

[36] Muhlberger E.. Filovirus replication and transcription *Future Virology.* 2007;2(2):205-215.

[37] Lei Y., Zhang J., Schiavon C.R., et al. SARS-CoV-2 Spike Protein Impairs Endothelial Function via Downregulation of ACE 2 *Circulation Research.* 2021;128:1323-1326.

[38] Zaidi A.K., Dehgani-Mobaraki P.. The mechanisms of action of ivermectin against SARS-CoV-2. An extensive review *The Journal of Antibiotics.* 2022.;75(2):60-71.

[39] Caly L., Druce J.D., Catton M.G., Jans D.A., Wagstaff K.M.. The FDA-approved drug ivermectin inhibits the replication of SARS-CoV-2 in vitro *Antiviral Research.* 2020;178:104787.

[40] Kory P., Meduri , Gianfranco U., Varon J., Iglesias J., Marik P.E.. Review of the Emerging Evidence Demonstrating the Efficacy of Ivermectin in the Prophylaxis and Treatment of COVID-19 *American Journal of Therapeutics.* 2021;28:e299-e318.

[41] Marik P.E., Kory P.. Ivermectin, A Reanalysis of the Data *American Journal of Therapeutics.* 2021;28 (5):e579-e580.

[42] Bryant A., Lawrie T.A., Dowswell T., et al. Ivermectin for Prevention and Treatment of COVID-19 Infection: A Systematic Review, Meta-analysis, and Trial Sequential Analysis to Inform Clinical Guidelines *American Journal of Therapeutics.* 2021;28 (4):e434-e460.

[43] Santi A.D., Scheim D.E., McCullough P.A., Yagisawa M., Borody T.J.. Ivermectin: a multifaceted drug of Nobel prize-honoured distinction with indicated efficacy against a new global scourge, COVID-19 *New Microbes and New Infections.* 2021;43:100924.

[44] Hazan S., Dave S., Gunaratne A.W., et al. Effectiveness of ivermectin-based multidrug therapy in severely hypoxic, ambulatory COVID-19 patients *Frontiers in Medicine.* 2022. https://doi.org/10.2217/fmb-2022-0014.

[45] Biancatelli R.M.L.C., Berrill M., Marik P.E.. The antiviral properties of vitamin C *Expert Review of Anti-infective Therapy.* 2020;18 (2):99-101.

[46] Grant W.B., Lahore H., McDonnell S.L., et al. Evidence that Vitamin D Supplementation Could Reduce Risk of Influenza and COVID-19 Infections and Deaths *Nutrients.* 2020;12:988.

[47] Mercola J., Grant W.B., Wagner C.L.. Evidence Regarding Vitamin D and Risk of COVID-19 and Its Severity *Nutrients.* 2020;12:3361.

[48] Risch H.A.. 2021. personal communication.

[49] Collins R., Bowman L., Landray M., Peto R.. The Magic of Randomization versus the Myth of Real-World Evidence *New England Journal of Medicine.* 2020;382(7):674-678.

[50] Frieden T.R.. Evidence for Health Decision Making - Beyond Randomized, Controlled Trials *New England Journal of Medicine.* 2017;377:465-475.

[51] Deaton A., Cartwright N.. Understanding and misunderstanding randomized controlled trials *Social Science and Medicine.* 2018;210:2-21.

[52] New Coronavirus Pneumonia Chinese Center for Disease Control, Prevention . Analysis of Epidemiological Characteristics of New Coronavirus Pneumonia *Chinese Journal of Epidemiology.* 2020;41:145-151.

[53] Onder G., Rezza G., Brusaferro S.. Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy *Journal of the American Medical Association.* 2020;323:1775-1776.

[54] Verity R., Okell L.C., Dorigatti I., et al. Estimates of the severity of coronavirus disease 2019: A model-based analysis. *The Lancet Infectious Diseases.* 2020;20:669-677.

[55] Barda N., Riesel D., Akriv A., et al. Developing a COVID-19 mortality risk prediction model when individual-level data are not available *Nature Communications.* 2020;11:4439.

[56] CDC . Estimated COVID-19 Burden Accessed on 11/16/2021 from https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html.

[57] Reese H., Iuliano A.D., Patel N.N., et al. Estimated Incidence of Coronavirus Disease 2019 (COVID-19) Illness and Hospitalization-United States, February-September 2020 *Clinical Infectious Diseases.* 2021;72:e1010-e1017.

[58] Iuliano A.D., Chang H.H., Patel N.N., et al. Estimating underrecognized COVID-19 deaths, United States, March 2020-May 2021 using an excess mortality modelling approach *Lancet Region. Health - Americas.* 2021;1:100019.

[59] Risch H.A.. Early Outpatient Treatment of Symptomatic, High-Risk COVID-19 Patients that Should be Ramped-Up Immediately as Key to the Pandemic Crisis. *American Journal of Epidemiology.* 2020;189:1218-1226.

[60] Rudrapal M., Khairnar S.J., Jadhav A.G.. Drug Repurposing (DR): An Emerging Approach in Drug Discovery in *Drug Repurposing* (Badria F.A.. , ed.)(Rijeka)IntechOpen 2020.

[61] Leemis L., McQueston J.. Univariate Distribution Relationships *The American Statistician.* 2008;62:45-53.

[62] Leemis L.M., Luckett D.J., Powell A.G., Vermeer P.E.. Univariate Probability Distributions *Journal of Statistical Education.* 2012;20:1-11.

[63] Lopez-Blazquez F., Mino B. Salamanca. Binomial approximation to hypergeometric probabilities *Journal of Statistical Planning and Inference.* 2000;87:21-29.

[64] Reiczigel J.. Confidence intervals for the binomial parameter: some new considerations *Statistics in Medicine.* 2003;22:611-621.

[65] Park H., Leemis L.M.. Ensemble Confidence Intervals for Binomial Proportions *Statistics in Medicine.* 2019;38:3460-3475.

[66] Brown Lawrence D., Cai T. Tony, DasGupta Anirban. Interval Estimation for a Binomial Proportion *Statistical Science.* 2001;16 (2):101-133.

[67] Clopper C., Pearson E.S.. The use of confidence or fiducial limits illustrated in the case of the binomial *Biometrika.* 1934;26:404-413.

[68] Sterne T.E.. Some remarks on confidence or fiducial limits *Biometrika.* 1954;41:275-278.

[69] Crow E.L.. Confidence intervals for a proportion *Biometrika.*

1956;43:423-435.

[70] Blyth C.R., Still H.A.. Binomial Confidence Intervals *Journal of the American Statistical Association.* 1983;78 (381):108-116.

[71] Benjamin D.J., Berger J.O.. Three Recommendations for Improving the Use of p-Values *The American Statistician.* 2019;73:sup1:186-191.

[72] Vidgen B., Yasseri T.. P-Values: Misunderstood and Misused *Frontiers in Physics.* 2016;4(6):6pp.

[73] Jeffreys H.. *The Theory of Probability*. Oxford UK: Oxford University Press 1998.

[74] Maxima . Maxima, a Computer Algebra System. Version 5.41.0 http://maxima.sourceforge.net/ 2017.

[75] Risch H.A.. The author replies *American Journal of Epidemiology.* 2020;189:1444-1449.

[76] Hill A.B.. The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine.* 1965;58(5):295-300.

[77] Data Our World. Coronavirus Pandemic (COVID-19) Accessed on 11/02/2021 from https://ourworldindata.org/coronavirus.

[78] Fonseca S.N.S., Sousa A.Q., Wolkoff A.G., et al. Risk of hospitalization for COVID-19 outpatients treated with various drug regimens in Brazil: Comparative analysis *Travel Medicine and Infectious Disease.* 2020;38:101906.

[79] Poly T.N., Islam M.M., Yang H.C., et al. Obesity and Mortality Among Patients Diagnosed With COVID-19: A Systematic Review and Meta-Analysis *Frontiers in Medicine.* 2021;8:620044.

[80] Kompaniyets L., Goodman A.B., Belay B., et al. Body Mass Index and Risk for COVID-19-Related Hospitalization, Intensive Care Unit Admission, Invasive Mechanical Ventilation, and Death - United States, March-December 2020 *Morbidity and Mortality Weekly Report.* 2021;70(10):355-361.

[81] Battineni G., Sagaro G.G., Chintalapudi N., Amenta F., Tomassoni D., Tayebati S.K.. Impact of Obesity-Induced Inflammation on Cardiovascular Diseases (CVD) *International Journal of Molecular Sciences.* 2021;22(9):4798.

[82] Harbi M. Al, Kaabi N. Al, Nuaimi A. Al, et al. Clinical and laboratory characteristics of patients hospitalised with COVID-19: Clinical outcomes in Abu Dhabi, United Arab Emirates *BMC Infectious Diseases.* 2022;22(1):136.

[83] Ulhaq Z., Soraya G.. Interleukin-6 as a potential biomarker of COVID-19 progression *Medecine et Maladies Infectieuses.* 2020;50(4):382-383.

[84] Vahidy F.S., Drews A.L., Masud F.N., et al. Characteristics and outcomes of COVID-19 patients during initial peak and resurgence in the Houston metropolitan area *Journal of the American Medical Association.* 2020;324:998-1000.

[85] Palazzuoli A., Ruberto F., Ferrari G.M. De, et al. Inpatient mortality according to level of respiratory support received for severe acute respiratory syndrome coronavirus 2 (coronavirus disease 2019) infection: A prospective multicenter study *Critical Care Explorations.* 2020;2:e0220.

[86] Hatfill S.. The Intentional Destruction of the National Pandemic Plan *Journal of the American Physicians and Surgeons.* 2021;26:74-76.

[87] Jehi L., Ji X., Milinovich A., et al. Development and validation of a model for individualized prediction of hospitalization risk in 4,536 patients with COVID-19 *PLoS ONE.* 2020;15:e0237419.

[88] Dashti H., Roche E.C., Bates D.W., Mora S., Demler O.. SARS-2 simplified scores to estimate risk of hospitalization and death among patients with COVID-19 *Nature Scientific Reports.* 2021;11:4945.

[89] Siddiqi H.K., Libby P., Ridker P.M.. COVID-19 - A vascular disease *Trends in Cardiovascular Medicine.* 2021;31(1):1-5.

[90] Gavriilaki E., Anyfanti P., Gavriilaki M., Lazaridis A., Douma S., Gkaliagkousi E.. Endothelial Dysfunction in COVID-19: Lessons Learned from Coronaviruses *Current Hypertension Reports.* 2020;22(9):63.

[91] Hong Y.. On computing the distribution function for the Poisson binomial distribution *Computational Statistics and Data Analysis.* 2013;59:41-51.

[92] Mucchielli L.. Behind the French controversy over the medical treatment of Covid-19: The role of the drug industry *Journal of Sociology.* 2020;56:736-744.